ARI Research Note 96-46

The Effects of System Failure and Time Limitations on Problem-Solving Behavior and Performance

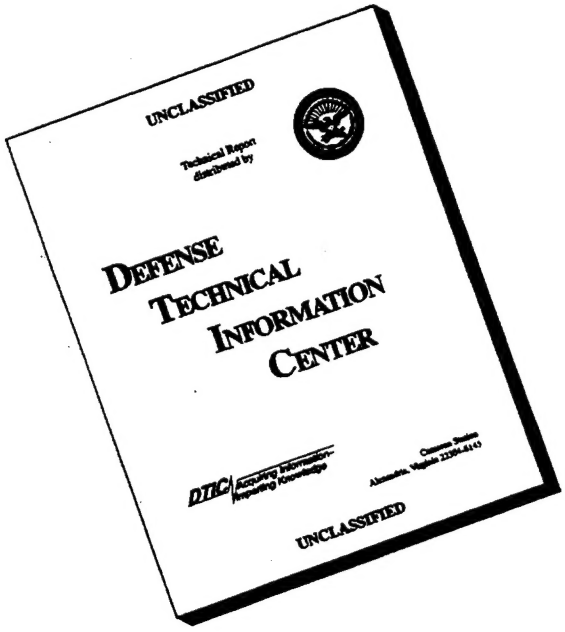
Bonnie J. Walker
Central State University

Research and Advanced Concepts Office
Michael Drillings, Acting Director

March 1996

United States Army
Research Institute for the Behavioral and Social Sciences

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

Central State University

Technical review by

Joseph Psotka

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1996, March	Final	June 1990-December 1993

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
The Effects of System Failure and Time Limitations on Problem-Solving Behavior and Performance	MDA903-90-C-0104
	5b. PROGRAM ELEMENT NUMBER 0601102A
6. AUTHOR(S)	5c. PROJECT NUMBER B74F
Bonnie Walker (Central State University)	5d. TASK NUMBER 3901
	5e. WORK UNIT NUMBER C02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Central State University 1400 Brush Row Road Wilberforce, OH 45384	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: PERI-BR 5001 Eisenhower Avenue Alexandria, VA 22333-5600	ARI
	11. MONITOR REPORT NUMBER Research Note 96-46

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

COR: Michael Drillings

14. ABSTRACT *(Maximum 200 words)*:

To explore the effects of system failure (data error) and time limitations on problem-solving behavior and performance, 12 experiments were conducted using two inferential reasoning tasks. In general, system failure and time limitations lead to significant decrements in performance. In addition, a protocol analysis of problem-solving behavior revealed under both normal and system failure was indicative of a lack of development of metacognitive strategies for working under unreliable conditions. It was recommended that system training under degraded modes of operation should include some provisions for imposing time limits for the completion of certain tasks.

15. SUBJECT TERMS

System failure Time limitations Problem-solving

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	77	

THE EFFECTS OF SYSTEM FAILURE AND TIME LIMITATIONS ON PROBLEM-SOLVING BEHAVIOR AND PERFORMANCE

EXECUTIVE SUMMARY

Research Requirement:

To explore the effects of system failure (data error) and time limitations on problem-solving behavior and performance in inferential reasoning tasks.

Procedure:

Based on the findings of earlier studies, 12 experiments, utilizing two new versions of basic inferential reasoning tasks and involving 424 subjects, were designed, conducted, and analyzed. Seven experiments were sequentially implemented using the Wason (1960) 2-4-6 rule discovery problem to initially replicate previous studies in which system failure in data feedback led to significant performance decrements and then to study the effects of imposed time limits on task completion. Two additional experiments, a protocol analysis and a training task again using the Wason problem, were conducted to provide a more in-depth view of problem-solving behavior under system failure conditions and to explore the use of an analogous reasoning task for training. Three experiments using the Kern (1982) artificial universe problem were conducted to study the effects of various levels of system failure in data feedback on overall performance.

Findings:

In general, system failure led to significant decrements in performance on both tasks. For subjects attempting to solve the artificial universe problem, an increase in system failure from low to high error rates dramatically decreased solving rates. The majority of experiments using the Wason problem also demonstrated a significant effect for system failure in decreased solving rates. The addition of imposed time limits in system failure conditions further decreased solving rates.

The protocol analysis of performance under normal and system failure conditions demonstrated that many subjects used "strong inference" (Platt, 1964) in attempting to solve the Wason problem. It was discovered that the majority of solvers in both the normal and system failure conditions considered several potential solutions simultaneously and systematically eliminated them by using test feedback. The strategy was effective under the system failure conditions if critical tests were repeated. It appeared that nonsolvers failed to consider the relationship between data feedback and all potential solutions.

The use of an analogous but more realistic scenario for training subjects to eliminate ideas and recognize relevant data was not successful in increasing solution rates in the Wason task. Instructions to disconfirm ideas and repeat tests also did not change subjects' problem-solving heuristics significantly.

Utilization of Findings:

The poor performance of many subjects under system failure conditions in such simple laboratory tasks is indicative of a lack of knowledge of effective metacognitive strategies for working under unreliable conditions. Complex systems with varying rates of reliability for each component have numerous sources of data error. Developing effective strategies for coping with data error is, thus, crucial for efficient operation of complex, multitask systems. It is felt that the development of coping strategies should occur during system training through regular implementation of degraded modes of operation. Steps must be taken to ensure that operators learn what procedures should be followed to function efficiently in degraded mode. Furthermore, training under degraded mode should also include some provision for imposing time limits on completion of specific tasks.

It is also felt that individual differences in problem-solving approaches should be taken into account when implementing training programs. Increased emphasis should be placed on the standardized assessment of analytical skills of prospective personnel for complex system operation. Such assessment could be used to tailor training to particular needs, decrease the amount of training necessary, and lower the dropout rate from highly technical training programs.

THE EFFECTS OF SYSTEM FAILURE AND TIME LIMITATIONS ON PROBLEM-SOLVING BEHAVIOR AND PERFORMANCE

CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF FIGURES (Continued)

THE EFFECTS OF SYSTEM FAILURE AND TIME LIMITATIONS ON PROBLEM-SOLVING BEHAVIOR AND PERFORMANCE

Introduction

Most experimental tasks designed to study problem-solving processes, such as how hypotheses are discovered and tested (i.e., Wason, 1960; Mynatt, Doherty & Tweney, 1977, 1978; Tweney et al., 1980), have provided subjects with an ideal, error-free testing environment. In reality, however, completely error-free data are rarely available and theoretical and practical inferences are routinely made based on data which contain varying degrees of error (false negative and/or false positive feedback). For instance, false alarms (false positive feedback) can be triggered by a momentary failure (power surges, excessive heat, vibration) of an automated system (smoke alarms, burglar alarms, and automobile theft alarms). The occurrence of false alarms has also been documented in complex task environments, such as medical care units (Kerr, 1985), automobiles (Caelli and Porter, 1980), aircraft (Billings, 1991), and nuclear power plants (Kantowitz, 1977). The possibility of system failures produces ecological unreliability (Brunswick, 1956) in both data generation and evaluation. In turn, such unreliability may have serious adverse effects on both the problem-solving behavior and task performance of an individual system operator. The effects on problem-solving behavior may include the strengthening of the tendency to only look for feedback which conforms to what is expected to occur, failing to replicate crucial tests, and failing to utilize relevant, but disconfirmatory, data. The inability to apply appropriate problem-solving heuristics when faced with an unreliable system may, thus, result in degraded task performance.

Several basic research studies using a variety of laboratory tasks have attempted to assess how false positive and/or false negative feedback generally affect problem-solving heuristics and ultimately affect task performance. For example, the "cry-wolf" phenomenon has been found in numerous studies (Breznitz, 1983; Pate-Cornell, 1986; Bliss, 1993) of responses to false alarms. The effects of false alarms have been found to range from complete response cessation (Pate-Cornell, 1986) to various levels of degraded response (Breznitz, 1983).

The "bias to confirm" an idea is also a well-documented phenomenon (Wason, 1960; Tweney et al., 1981; Gorman & Gorman, 1984; Walker, 1985, 1987; Doherty and Tweney, 1988; Walker & Harper, 1989, 1990) which occurs when an individual is selecting data to test a particular belief or idea. Such a bias may be particularly useful during hypothesis development and/or to establish data reliability (Mynatt, Doherty & Tweney, 1977, 1978; Tweney, Doherty & Mynatt, 1981; Klayman & Ha, 1987, Tweney, 1985;

Tukey, 1986). However, complete dependence on such a heuristic may lead to erroneous conclusions, especially when used under unreliable conditions.

In a study designed to assess the interactive effect of confirmation bias and false feedback in a group problem-solving task, Gorman (1986) found severe disruption of task performance under false feedback conditions. In addition, Gorman noted that most subjects assumed that data which refuted their ideas was false and either ignored it or showed a preference for replicating such trials. Kern (1982), as well as Doherty and Tweney (1988), used an artificial universe task to study the effects of actual, rather than possible, false feedback on performance and confirmation bias. Subjects launched imaginary creatures, dependent on moisture for survival, from a spaceship to the surface of an unexplored planet to discover a survival boundary. The planet surface's moisture content varied across the area and feedback concerning survival was given after every launch. In some conditions, subjects were warned that the feedback data might not always be correct and were provided with the opportunity to check a limited number of launch results. Results of both the Kern (1982) and Doherty and Tweney (1988) studies demonstrated task performance decrements in the false feedback conditions. Also, as Gorman (1986) had found, if the feedback did not support a subject's current belief about the location of the boundary line, disconfirming data was usually ignored as error or only disconfirming test outcomes were replicated.

Doherty and Tweney (1988) also explored the effects of system failure on inference and prediction using a multiple cue probability learning task (see also York, Doherty, & Kamouri, 1987). Subjects were given two or more numerical values and asked to predict another value. Following the prediction, the subjects were told the correct value. Under system failure conditions, either the initial values or the feedback about the correct values was sometimes wrong. Doherty and Tweney (1988) reported that when the task was kept simple, system failure had no significant effect on performance. However, when the complexity of the task increased, successful task performance became much more difficult.

Utilizing the Wason (1960) 2-4-6 rule discovery task, Walker (1987) differentiated between how only the knowledge that system failure might occur and actual system failure (both false positive and negative feedback) affected problem-solving heuristics and task performance. Subjects were given an initial number sequence, "2-4-6", and asked to discover a general number-sequencing rule by testing additional sequences. The feedback from the test results was used to eventually declare a rule. Walker found that problem-solving heuristics, including confirmation bias, were relatively unaffected by whether or not

subjects knew system failure might occur. However, solving rates decreased significantly when subjects were warned that a system failure might occur and no error was actually present, since relevant data which disconfirmed a currently held belief was ignored as data error. In addition, actual system failure increased the number of tests conducted, as well as the number of replications, which further degraded task performance.

Walker and Harper (1989) extended the Walker (1987) methodology to include three levels of actual system failure and used a scientifically sophisticated sample of active researchers as subjects, rather than the usual undergraduate sample. It had been hypothesized that active researchers, trained in scientific procedures and experienced to varying degrees with actual unreliable data, would demonstrate development of appropriate heuristics for dealing with such problems. However, the researchers showed a strong bias to confirm their ideas and task performance was significantly undermined by system failure. In a follow-up study, Walker and Harper (1990) used engineers and non-engineers as subjects and modified the task to compare restricted and unrestricted potential rule choices. The results revealed that the engineers were less likely to be confused by system failure in the restricted condition and more likely to utilize the results of test strategies (test replication and disconfirmation) to check for data error and rule out competing ideas. In the unrestricted conditions, both with and without system failure, the number of overall tests conducted decreased and task performance was poor. Walker and Harper (1990) noted that many of the engineers and non-engineers in these conditions expressed concern over completing the task quickly and returning to their work assignments, which may have indicated that perceived time constraints also affected performance.

The results of these basic studies are indicative of the effects system reliability can have on a human operator's performance, as well as the effects of environmental stressors and individual differences in the use of problem-solving heuristics on inferential tasks. The current study was designed to assess the effects of both time limitations and system failure on problem-solving heuristics and task performance. The study utilized two experimental paradigms--the Wason (1960) 2-4-6 rule discovery problem and the Kern (1982) "Tribbles" task. It was initially hypothesized, based on the results of the Walker and Harper (1990) study, that imposed time limitations and the interjection of system failure into feedback data would decrease solving efficiency on the Wason (1960) task. It was also hypothesized, based on the results of the Kern (1982) and Doherty and Tweney (1988) studies, that system failure would seriously disrupt solving efficiency on the "Tribbles" task.

During the current study a total of 12 experiments were completed involving 424 subjects from two college undergraduate

subject pools. The first six experiments were conducted at Central State University, Wilberforce, Ohio. The last six experiments were conducted at the University of Central Florida, Orlando, Florida. No comparisons between subject pools were included in the analyses.

Phase I--"Wason 2-4-6" Replication Attempts

Experiment 1. No Error vs. Error Conditions

Rationale. The first experiment was conducted to assess the replicability of the Walker and Harper (1990) methodology.

Subjects. Twenty (20) Central State University undergraduate students participated in the study. All subjects were paid $5.00 for their participation. (Note: Two subjects did not complete the task and the data was omitted from the final analysis.)

Procedure. The experiment was designed to compare a 15% data error condition to a no error condition, using only the restricted versions of Walker and Harper's (1990) original software. Ten subjects were randomly assigned to either a No Error or Error condition. Each program compared a three-digit keyboard entry (e.g., 1, 3, 5) to a general number-sequencing rule, "three ascending numbers". For the Error condition, a subroutine was randomly activated for 15% of the data entries. The subroutine reversed the computer response to a data entry so that a sequence that actually fit the "ascending numbers" rule was responded to as not fitting and vice-versa.

Subjects first read a hard-copy of the task instructions along with the experimenter and were given an opportunity to ask for further clarification. An abbreviated version of the task instructions was then brought up on the computer monitor and the subjects were asked to review them. The display included the sequence "2, 4, 6" as an example that fit the rule and a highlighted warning about the possibility of incorrect computer responses in the Error condition.

At the top of the next screen (the testing screen) the following column headings were displayed: "My Idea"; "Test Sequence"; and "Fits (Y/N)". Under the "Test Sequence" and "Fits (Y/N)" columns, the sample sequence, "2-4-6", and "Yes" were displayed. The bottom screen display queried the subject to respond to the question, "Which of the following rules do you believe fits the number sequence above?" by selecting one of five possibilities: "(1) Even numbers, (2) Numbers ascending by 2, (3) Numbers less than 10, (4) Ascending numbers or (5) Other."

Following entry of a number, a new screen queried subjects about testing the rule they had selected and entering a three-digit sequence. After each entry, but before the program's response to the test, subjects were instructed to indicate whether or not they thought their test would fit the rule. For instance, if a subject's first rule selection was "Even numbers"

and the sequence "8, 10, 12" was entered, the screen displayed the question: "Do you think 8, 10, 12 will fit the rule?" Subjects indicated their expected test outcomes by entering "Y" (yes), "N" (no), or "U" (unsure).

Following entry of the expected outcome, the program displayed one of two responses to the number-sequence test: (1) "That sequence fits the rule" or (2) "That sequence does not fit the rule". All attempted tests and results were displayed at the bottom of the screen for continuous review while conducting subsequent tests.

When subjects were ready to announce a rule, they exited the testing screen and made their selection from a duplicate list of possible rules. They were then told whether or not the rule announcement was correct.

Results. Initial scanning of subject protocols indicated that one subject in each of the two conditions exited the task without performing any tests and the data was dropped from the final analyses. In the No Error condition, 1 out of 9 subjects (11.1%) was able to solve the task, compared to 5 out of 9 subjects (55.6%) who were able to solve in the Error condition (see Figure 1). The difference between the solving rate proportions was significant ($z = 3.21$, $p < .01$, two-tailed).

Five measures were used to indicate problem-solving heuristics--total attempted tests, test repetition, and test outcome expectations ("Yes", "No", and "Unsure"). The mean number of tests conducted in the No Error condition was 4.2, compared to 5.9 tests in the Error condition. The difference in the number of tests conducted between the conditions was not significant ($t(16) = 1.289$, NS).

The mean percentage of total tests that were repeated sequences in the No Error condition was 33.3% compared to 32.4% in the Error condition. The difference in the mean percentages of repeated tests between the conditions was not significant ($t(16) = .235$, NS). There was also no significant difference between the conditions in the proportions of subjects who repeated tests ($z = -.676$, NS, two-tailed). In the No Error condition, 4 out of 9 subjects (44.4%) repeated tests, compared to 5 out of 9 subjects (55.6%) in the Error condition.

There was no significant difference between conditions in the mean percentages of total tests attempted that subjects' expected would fit the computer's rule ($t(16) = 1.146$, NS). Subjects in the No Error condition expected 79.2% of their attempted tests to fit the rule, while subjects in the Error condition expected 58.2% to fit. The difference between conditions in the percentages of total tests attempted that subjects expected would not fit the computer's rule was also not significant ($t(16) =$

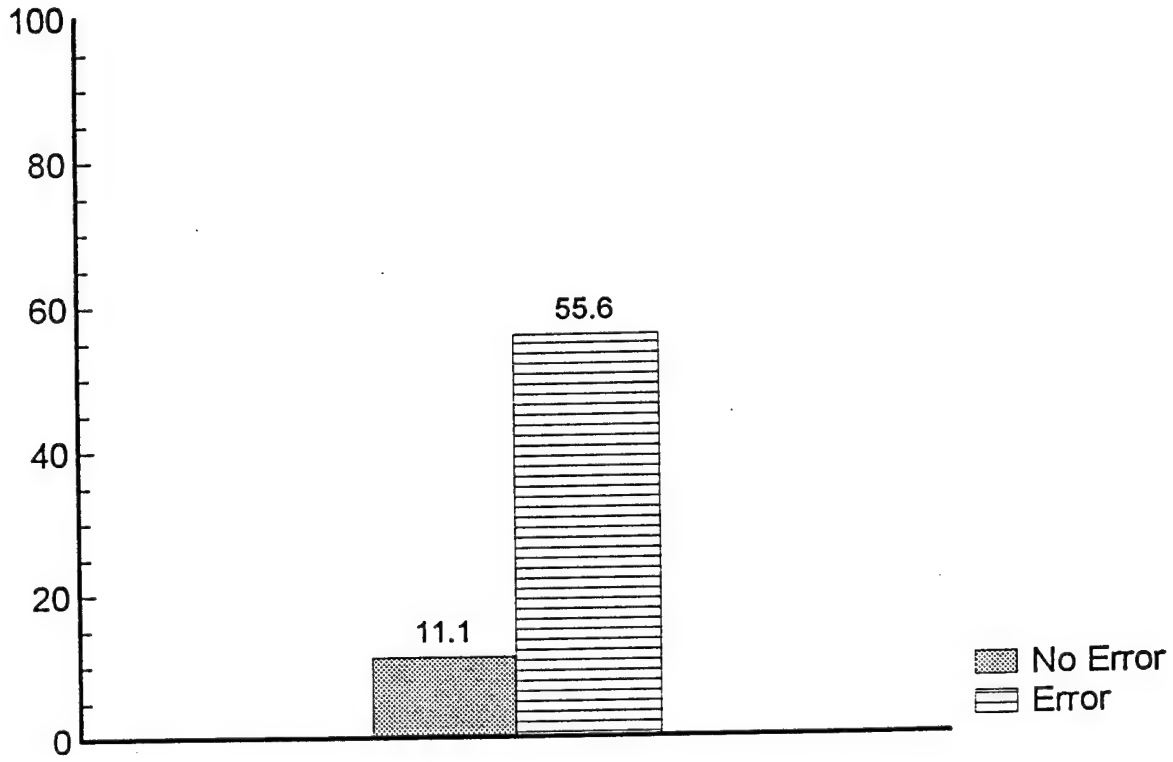


Figure 1. Experiment 1--No Error vs. Error Percentage of Solvers

'.047, NS). Subjects in the No Error condition indicated that they did not expect 7.1% of their attempted test tests to fit the rule, while subjects in the Error condition expected 7.5% not to fit. Similarly, the difference between conditions in the percentages of "Unsure" responses was not significant ($t(16)$ = 1.665, NS). Subjects in the No Error condition indicated they were unsure of 13.6% of the attempted test outcomes while subjects in the Error condition indicated they were unsure of 1.2% of the test outcomes.

To determine whether the actual test outcomes confirmed or disconfirmed subjects' expectations, expected and actual outcomes for each test were combined. Matching combinations were classified as confirmatory and mismatching combinations were classified as disconfirmatory. No significant difference was found between the mean percentages of confirmatory test outcomes in the No Error condition (82.9%) and the Error condition (73.4%) ($t(16)$ = .994, NS). A significant difference was found between the mean percentages of disconfirmatory test outcomes in the No Error condition (3.4%) and the Error condition (25.4%) ($t(16)$ = 3.847, $p<.001$).

Discussion. While there was a significant difference between conditions in the proportion of solvers to non-solvers, it was not in the expected direction. The percentage of solvers (11.1%) in the No Error condition was much lower than the Walker and Harper (1990) findings in which 53.3% of the subjects solved the task. However, the percentage of solvers (55.6%) in the Error condition was slightly greater than the percentage of solvers (40.0%) in Walker and Harper's error condition.

An analysis of subject protocols for both conditions indicated that many non-solvers were exiting the task prematurely. While the task can be solved using a low number of trials to rule out competing hypotheses (a disconfirmatory strategy), the majority of non-solvers attempted only two or three confirmatory trials and stated the rule. The one subject who solved the task in the No Error condition used two disconfirmatory trials to reach the correct solution. In the Error condition, solvers used a higher mean number of trials (7.4), compared to non-solvers (4.0) and received a higher percentage of disconfirmatory feedback.

Experiment 2. No Error vs. Error; Imposed Minimum Times

Rationale. The purpose of the second experiment was an attempt to control subjects' early withdrawal from the task by requiring a minimum time for participation.

Subjects. Thirty subjects (26 Central State University undergraduate students, as well as 4 high school students enrolled in an Upward Bound summer program) participated in the study. All subjects were paid $5.00 for their participation.

Procedure. The experiment utilized the same procedure as Experiment 1. Fifteen subjects were randomly assigned to one of two conditions: (1) No Error; and (2) Error. However, unlike Experiment I, all subjects were **required** to spend 20 minutes working on the task before making a rule announcement.

Results. In the No Error condition, 6 out of 15 subjects (40.0%) were able to solve the task, compared to 7 out of 15 subjects (46.6%) in the Error condition (see Figure 2). The difference in solving rate proportions between the conditions was not significant (z = .47, *NS*, two-tailed).

The mean number of tests conducted by subjects in the No Error condition was 11.3, compared to 12.3 trials for the Error condition. The difference in the mean number of total attempted tests between the conditions was not significant ($t(28)$ = .623, NS).

The mean percentage of total tests that were repeated sequences in the No Error condition was 25.6% compared to 26.6% in the Error condition. The difference in the mean percentages of repeated tests between the conditions was not significant ($t(28)$ = .439, NS). There was a significant difference between conditions in the proportions of subjects who repeated tests (z = 3.35, $p<.01$, two-tailed) (see Figure 3). In the No Error condition, 13 out of 15 (86.7%) subjects repeated tests compared to 10 out of 15 (66.7%) subjects in the Error condition.

There was no significant difference between conditions in the mean percentages of total tests attempted that subjects' expected would fit the computer's rule ($t(28)$ = .338, NS). Subjects in the No Error condition expected 83.5% of their attempted tests to fit the rule, while subjects in the Error condition expected 79.8% to fit. The difference between conditions in the mean percentages of total tests attempted that subjects did not expect to fit the computer's rule was also not significant ($t(28)$ = .859, NS). Subjects in the No Error condition did not expect 4.8% of their total attempted tests to fit the rule, while subjects in the Error condition did not expect 9.2% to fit. Similarly, the difference between conditions in the mean

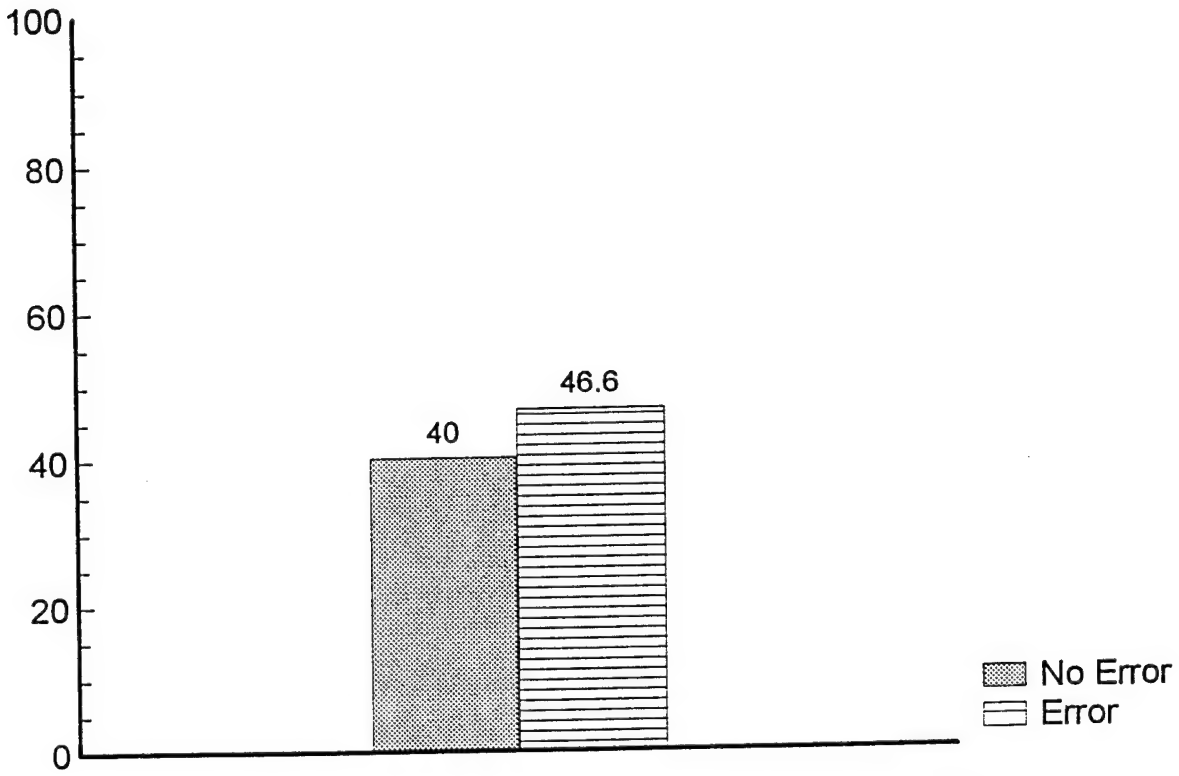


Figure 2. Experiment 2--Imposed Minimum Time Limit
Percentage of Solvers

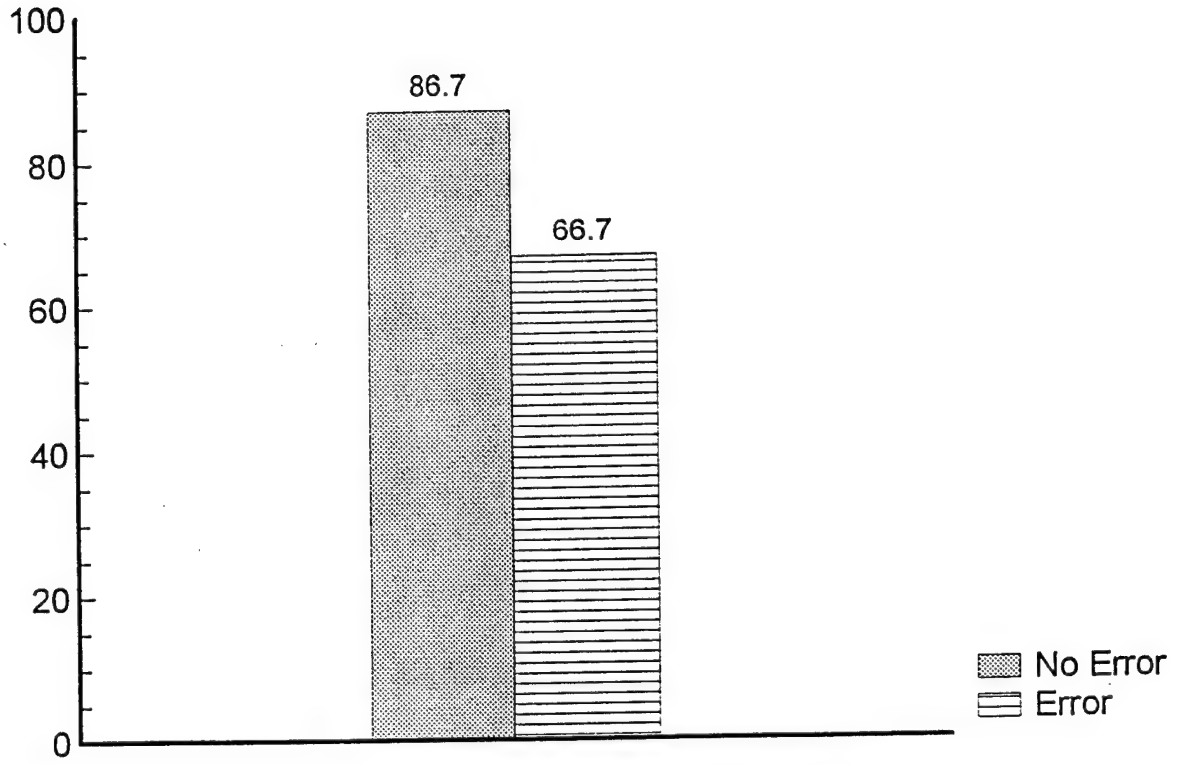


Figure 3. Experiment 2--Minimum Time Limit
Percentage of Subjects Repeating Tests

percentages of "Unsure" responses was not significant ($t(28)$ = .085, NS). Subjects in the No Error condition indicated they were unsure of 11.7% of the attempted test outcomes while subjects in the Error condition were unsure of 12.6% of the test outcomes.

No significant difference was found between the mean percentages of confirmatory test outcomes in the No Error condition (70.2%) and the Error condition (60.6%) ($t(28)$ = .971, NS). No significant difference was found between the mean percentages of disconfirmatory test outcomes in the No Error condition (15.3%) and the Error condition (26.8%) ($t(28)$ = 1.854, NS).

Discussion. The percentage of solvers (40.0%) in the No Error condition was somewhat lower than the Walker and Harper (1990) findings (53.3%), but considerably higher than had been found in Experiment 1 (11.1%). The percentage of solvers (46.6%) in the Error condition was again higher than Walker and Harper's results (40.0%), but lower than had been found in Experiment 1 (55.6%). The results, therefore, indicated that, overall, the minimum time requirement appeared to stabilize subjects' performance in both conditions by preventing early task withdrawal.

Furthermore, subjects given error were equally likely to repeat tests even though they received more disconfirmation than no error subjects indicating that the use of the repetition heuristic might not be related to disconfirmatory test outcomes.

Phase I--General Discussion. By requiring subjects to spend 20 minutes engaged in discovering the task solution, task performance was stabilized. Therefore, it was felt that investitagation into how time limitations affected performance could be investigated.

Phase IIA--Time Limitation Effects (Wason 2-4-6 Paradigm)

Experiment 3A. Imposed Time Limits; No Error Vs. Error

Rationale. The purpose of the third experiment was to introduce an artificial time constraint to the task to determine if increasing subjects' perceived stress under time pressure affected task performance.

Subjects. Forty subjects (39 Central State University undergraduate students and 1 faculty member) participated in the study. All undergraduate subjects were paid $5.00 for their participation.

Procedure. The experiment utilized the same procedure as Experiments 1 and 2 except that the task was conducted under Timed (10 and 20-minute limits) conditions crossed with No Error and Error conditions. Ten subjects were randomly assigned to one of four conditions: (1) 10-Minute No Error; (2) 10-Minute Error; (3) 20-Minute No Error; or (4) 20-Minute Error. For all conditions, a mechanical timer was placed on top of the video monitor and set to the appropriate time (10 or 20 minutes) after each subject completed reading the task instructions.

Results. In both the 10-Minute and 20-Minute No Error conditions, 5 out of 10 subjects (50.0%) were able to solve the task. In the 10-Minute Error condition, 3 out of 10 (30.0%) subjects, compared to 2 out of 10 (20.0%) subjects in the 20-Minute Error condition, solved the task. (See Figure 4.) The difference in the number of solvers compared to non-solvers among the conditions was not significant ($\chi^2(3, N = 40) = 2.88$, NS).

The mean number of tests conducted by subjects in both the 10-Minute and 20-Minute No Error conditions was 8.1. The mean number of tests conducted by subjects in the 10-Minute Error condition was 7.7, compared to 11.5 tests for the 20-Minute Error condition. A two-way ANOVA revealed no significant main effects for time or error among the conditions in the number of tests conducted ($F(1,36) = .957$, NS; $F(1,36) = 1.535$, NS).

The mean percentage of total attempted tests that were repeated in the 10-Minute No Error condition was 30.3% compared to 13.8% of the total attempted tests in the 20-Minute No Error condition. The mean percentage of total attempted tests that were repeated in the 10-Minute Error condition was 18.2%, compared to 27.1% of the total attempted tests in the 20-Minute Error condition. A two-way ANOVA revealed no significant main effects for time or error in the percentages of repeated tests conducted ($F(1,36) = 1.653$, NS; $F(1,36) = .031$, NS). However, there was a significant effect of condition on the number of individuals who repeated tests compared to those that did not

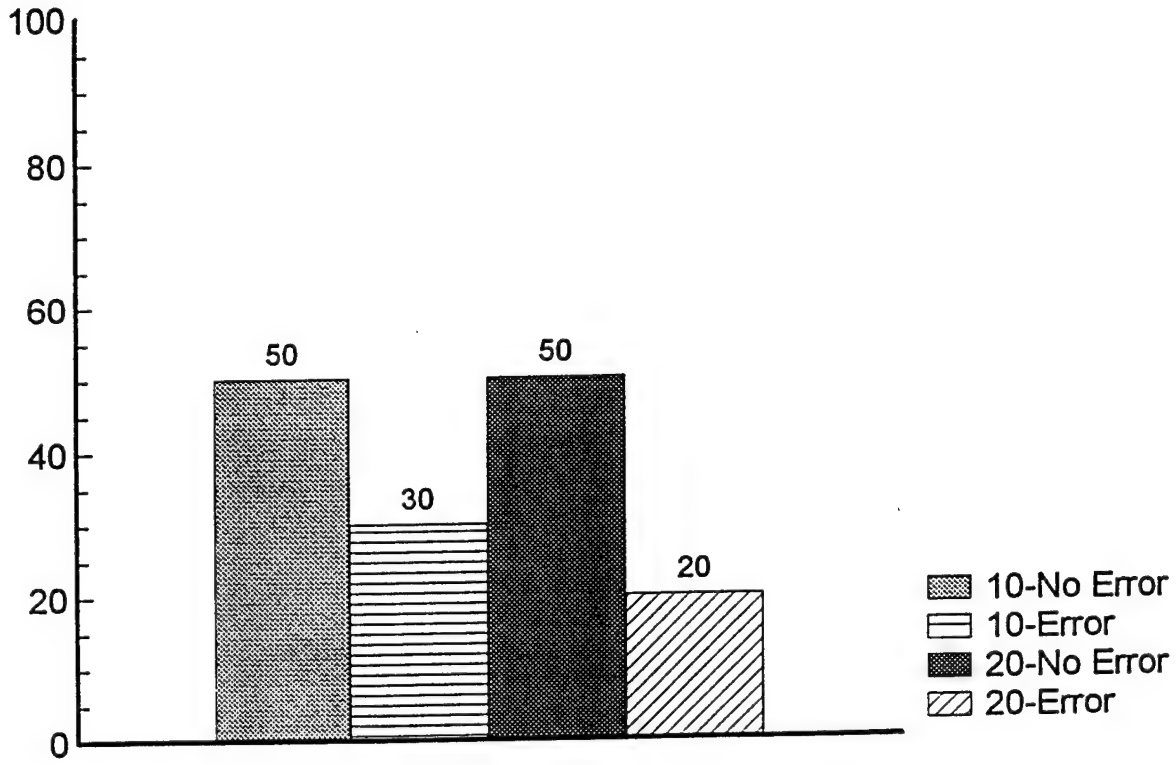


Figure 4. Experiment 3A--Imposed Time Limits
Percentage of Solvers

($\chi^2(3, N = 40) = 4.9, p<.05$). In the 10-Minute No Error condition, 4 out of 10 subjects (40.0%) repeated tests, compared to 3 out of 10 subjects (30.0%) in the 20-Minute No Error condition. In the 10-Minute Error condition, 6 out of 10 subjects (60.0%) repeated tests, compared to 7 out of 10 subjects (70.0%) in the 20-Minute Error condition. (See Figure 5.)

A two-way ANOVA indicated no significant main effects for time or error in the percentages of total tests attempted that subjects' expected would fit the computer's rule ($F(1,36) = .169$, NS; $F(1,36)) = .009$). Subjects in the 10-Minute No Error condition expected 84.0% of their attempted tests to fit the rule, while subjects in the 20-Minute No Error condition expected 85.2% to fit. Subjects in the 10-Minute Error condition expected 89.6% of their attempted tests to fit the rule, while subjects in the 20-Minute Error condition expected 85.2% to fit. The percentages of total tests attempted that subjects did not expect to fit the computer's rule also showed no significant main effects for time or error ($F(1,36) = 1.304$, NS; $F(1,36) = .598$, NS). Subjects in the 10-Minute No Error condition did not expect 9.9% of their attempted tests to fit the rule, while subjects in the 20-Minute No Error condition did not expect 11.1% to fit. Subjects in the 10-Minute Error condition did not expect 1.3% of their attempted tests to fit the rule, while subjects in the 20-Minute Error condition did not expect 7.0% to fit. Similarly, no significant main effects for time or error were found for the percentages of "Unsure" responses ($F(1,36) = .645$, NS; $F(1,36) = .631$, NS). Subjects in the 10 and 20-Minute No Error conditions indicated they were unsure of 6.2% and 3.7% of the test outcomes, while subjects in the 10 and 20-Minute Error conditions indicated they were unsure of 9.1% and 9.6% of the test outcomes.

There was a significant main effect for error, but not for time, found among the mean percentages of confirmatory test outcomes ($F(1,36) = 13.247$, $p = .001$; $F(1,36) = .002$, NS). In the 10-Minute No Error condition, 79.0% of the test outcomes were confirmatory, compared to 77.8% in the 20-Minute No Error condition. In the 10-Minute Error condition, 59.7% of the test outcomes were confirmatory, compared to 61.7% in the 20-Minute Error condition. There was also a significant main effect found for error, but not for time, among the mean percentages of disconfirmatory test outcomes ($F(1,36) = 4.919$, $p = .033$; $F(1,36) = 1.578$, NS). In the 10-Minute No Error condition, 14.8% of the test outcomes were disconfirmatory, compared to 18.5% in the 20-Minute No Error condition. In the 10-Minute Error condition, 31.2% of the test outcomes were disconfirmatory, compared to 30.4% in the 20-Minute Error condition.

Discussion. Though the comparison of solving rates among the conditions was not statistically significant, the results were in the expected direction. The percentages of solvers (50.0%) in

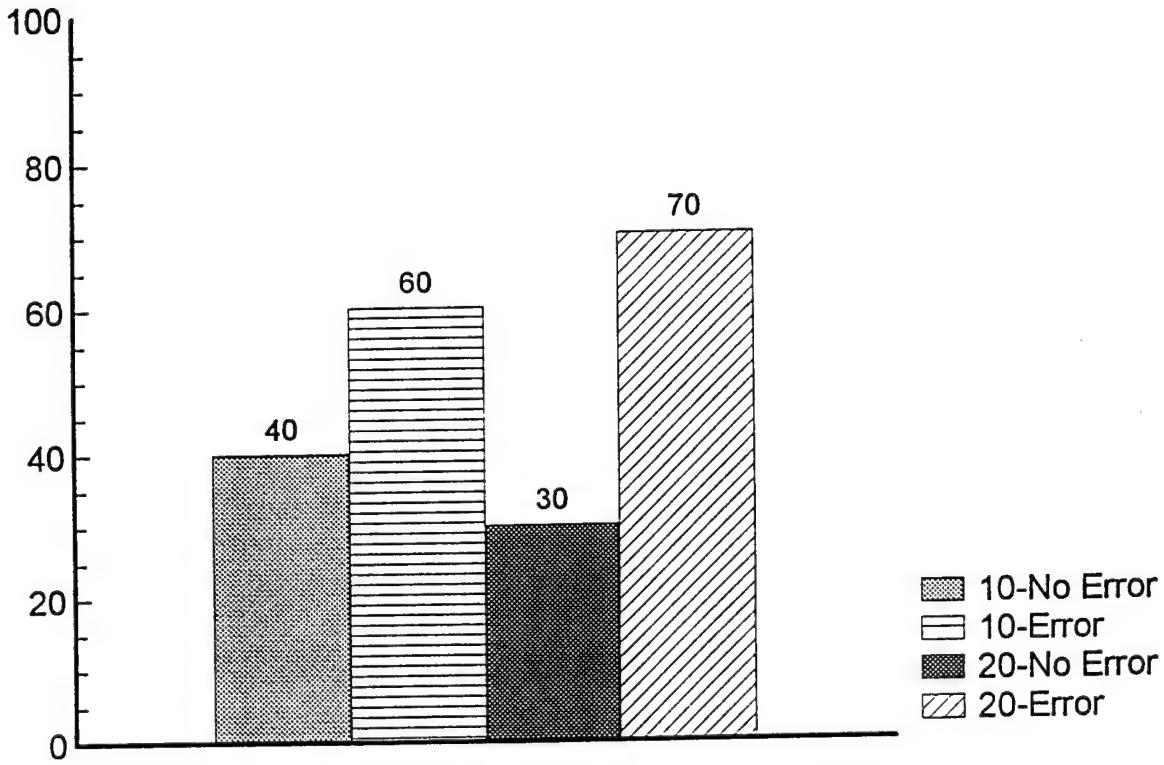


Figure 5. Experiment 3A--Imposed Time Limits
Percentage of Subjects Repeating Tests

both the 10 and 20-Minute No Error conditions were similar to Walker and Harper's (1990) findings in which 53.3% of the subjects solved the task and higher than the rate found in Experiment 2 (40.0%), indicating time constraints did not affect performance under normal conditions. The percentages of solvers (30.0% and 20.0%) in the 10 and 20-Minute Error conditions was lower than both the percentage of solvers (40.0%) in Walker and Harper's error condition and the percentages found in Experiments 1 and 2 (55.6% and 46.6%). Thus, it appeared that time constraints did disrupt task performance under system failure conditions.

In addition, a significant number of subjects repeated tests in the Error conditions which might be indicative of a change in problem-solving heuristics to check for erroneous feedback. There was also a significant decrease in the amount of confirmatory feedback subjects received in the error conditions and a significant increase in the amount of disconfirmatory feedback. It was felt that such differences in the kind of feedback received might have triggered the repetition of tests. Why test repetition did not improve solving rates in the error conditions is not known, though subjects seemed to be unable to effectively use the information from repeated tests to rule out competing hypotheses.

Experiment 4A--Decreased Time; No Error Vs. Error

Rationale. The fourth experiment was designed to compare performance under a decreased time constraint for both no error and error conditions.

Subjects. Forty-eight (48) Central State University undergraduate students participated in the study. All subjects were paid $5.00 for their participation.

Procedure. The experiment utilized the same procedure as the Experiment 3 except that the task was conducted under Timed (8 and 20-minute limits) conditions crossed with No Error and Error conditions. Twelve subjects were randomly assigned to one of the four experimental conditions: (1) No Error--8 minutes; (2) Error--8 minutes; (3) No Error--20 minutes; and (4) Error--20 minutes. A mechanical timer was placed on top of the video monitor and set to the appropriate time (8 or 20 minutes) after each subject completed reading the task instructions.

Results. Three out of 12 subjects (25.0%) in the 8-Minute No Error condition were able to solve the task, compared to all 12 subjects (100.0%) in the 20-Minute No Error condition. In both the 8-Minute and 20-Minute Error conditions, 5 out of 12 subjects (41.7%) were able to solve the task.

Due to the high solving rate found in the 20-Minute No Error condition, an analysis of subject protocols and a breakdown of experimenter/subject assignment was conducted. It was discovered that the internal validity of the experiment was confounded by both subject and experimenter biases and, thus, no further comparisons were deemed appropriate.

Discussion. It was apparent from reanalysis of the results of the study, that the program had to be redesigned so that correct task solutions would be randomly generated to avoid the problem of subjects sharing the correct outcome with their peers. In addition, following several discussions with both experimenters, it was revealed that many subjects required extra assistance in working through the task because they had difficulty understanding the concept of number sequences, as well as difficulty understanding the task instructions. Thus, it was felt that subjects should be given a pre-test of their knowledge of number-sequencing to correct any conceptual flaws and that the on-screen instructions should be clarified to minimize experimenter assistance.

Experiment 5A.--Timed vs. Untimed; No Error vs. Error

Rationale. The purpose of the experiment was to assess the effects of system failure on problem-solving behavior under timed and untimed conditions. Based on several methodological problems encountered with the Experiment 4, the procedures of Experiment 5A were changed to include a mathematical skills pretest, random correct solutions, and a simplified version of the original task instructions.

Subjects. Forty-two (42) Central State University undergraduate students participated in the experiment. All subjects were paid $5.00 for their participation.

Procedure. While the same basic procedures were followed as in the first four experiments, several changes were initiated. In contrast to the earlier versions of the task in which there was only one correct task solution (ascending numbers), the new computer program randomly generated one of four correct general number-sequencing rules: (1) Even numbers, (2) Numbers ascending by 2, (3) Numbers less than 10, and (4) Ascending numbers. All subjects were given two mathematical pretests. The first test involved completion of a series of four sequencing rules and three, three-digit sequence examples. Errors in completing the four types of sequences were corrected and explained by the experimenter. The second pretest involved asking subjects to match number-sequence examples to number-sequence rules. It should be noted that the rules used as examples on both pretests corresponded with the four rules used in the experiment. All verbal, hard-copy, and on-screen instructions were simplified. To insure the distinctiveness of the various portions of all screens, each segment of the display was highlighted in different colors.

The task was conducted under Timed (eight minutes maximum) and Untimed conditions crossed with No Error and Error conditions. Ten subjects were randomly assigned to one of the two Error conditions, Timed and Untimed; 11 subjects were randomly assigned to one of the two No Error conditions, Timed and Untimed. For both timed conditions, a mechanical timer was placed on top of the video monitor and set for eight minutes after the subject completed reading the on-screen task instructions. No timer was used for the untimed conditions.

Results. Initial scanning of subject protocols indicated that five subjects in the Timed No Error condition and two subjects in the Untimed No Error condition prematurely exited the task after completing only two trials. Subsequently, the data was omitted from the analyses. In the Timed No Error condition, 3 out of 6 subjects (50.0%) solved the task compared to 6 out of 11 subjects (54.5%) in the Untimed No Error condition. Only 1 out of 10

subjects (10.0%) in the Timed Error condition was able to solve the task compared to 6 out of 8 subjects (75.0%) who were able to solve in the Untimed Error condition. (See Figure 6.) The differences in solving rates among the conditions was significant ($\chi^2(3, N = 35) = 8.295, p = .04$).

The mean number of tests conducted in the Timed No Error condition was 8.2, compared to 9.2 tests conducted in the Untimed No Error condition. The mean number of tests conducted in the Timed Error condition was 6.1, compared to 13.3 tests conducted in the Untimed Error condition. (See Figure 7.) A two-way ANOVA revealed a significant main effect for time in the number of tests conducted among the conditions ($F(1,31) = 3.99, p = .05$).

The mean percentage of total tests that were repeated in the Timed No Error condition was 21.8%, compared to 37.9% in the Untimed No Error condition. The mean percentage of total tests that were repeated in the Timed Error condition was 28.6%, compared to 12.2% in the Untimed Error condition. A two-way ANOVA indicated no significant main effects for time or error in the percentage of repeated tests conducted among the conditions ($F(1,31) = .525$, NS; $F(1,31) = .667$, NS). There was also no significant effect of condition on the number of individuals who repeated tests compared to those who did not repeat tests ($\chi^2(3, N = 35) = 1.554$, NS). Five out of 6 subjects (83.3%) in the Timed No Error condition repeated tests, compared to 6 out of 11 subjects (54.5%) in the Untimed No Error condition. Seven out of 10 subjects (70.0%) in the Timed Error condition repeated tests, compared to 5 out of 8 subjects (62.5%) in the Untimed Error condition.

A two-way ANOVA indicated no significant main effects for time or error in the percentages of total tests attempted that subjects' expected would fit the computer's rule ($F(1,31) = .001$, NS; $F(1,31) = .326$, NS). Subjects in the Timed No Error condition indicated that they expected 76.1% of their total attempted tests to fit the rule, while subjects in the Untimed No Error condition indicated that they expected 77.2% to fit. Subjects in the Timed Error condition expected 88.3% of their total attempted tests to fit the rule, while subjects in the Untimed Error condition expected 88.0% to fit.

A significant interaction between time and error was found among the percentages of total tests attempted that subjects did not expect to fit the computer's rule ($F(1,31) = 5.022, p = .032$). Subjects in the Timed No Error condition indicated that they did not expect 7.2% of their total tests to fit the rule, while subjects in the Untimed No Error condition indicated that they did not expect any (0.0%) of their total attempted tests to not fit the rule. Subjects in the Timed Error condition did not expect 1.7% to fit, while subjects in the Untimed Error condition did not expect 5.8% to fit. (See Figure 8.)

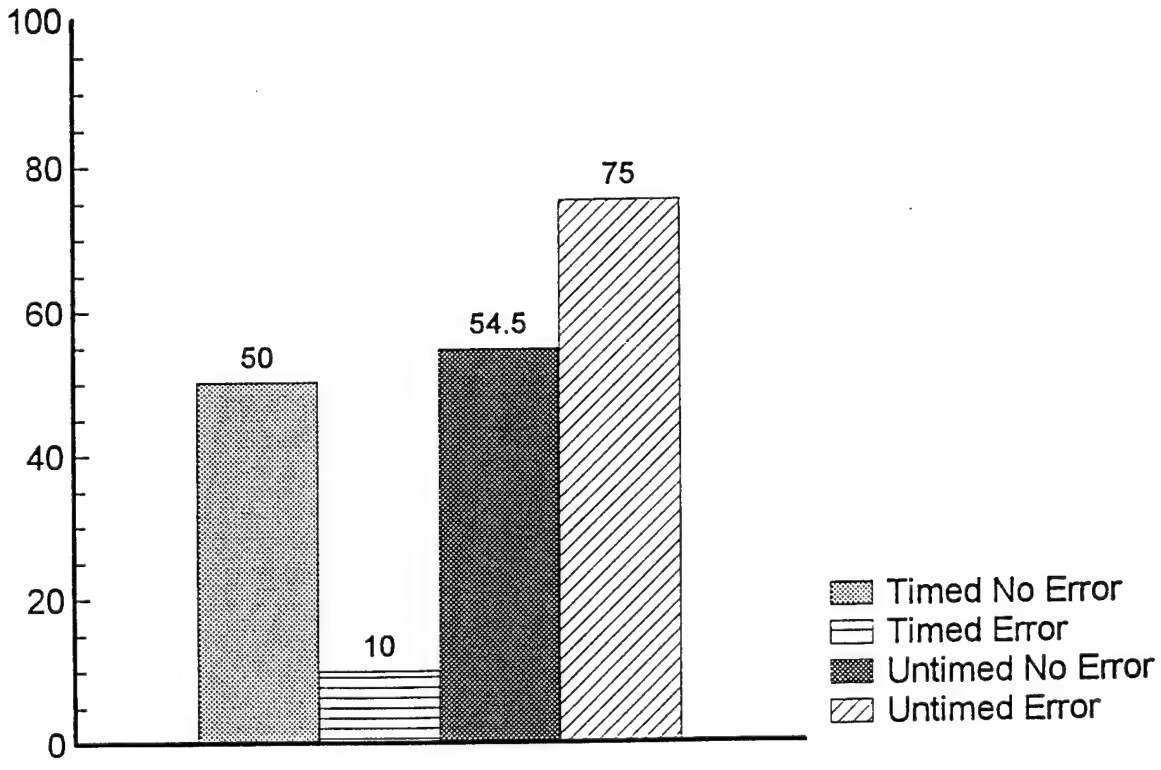


Figure 6. Experiment 5A--Timed vs. Untimed Percentage of Solvers

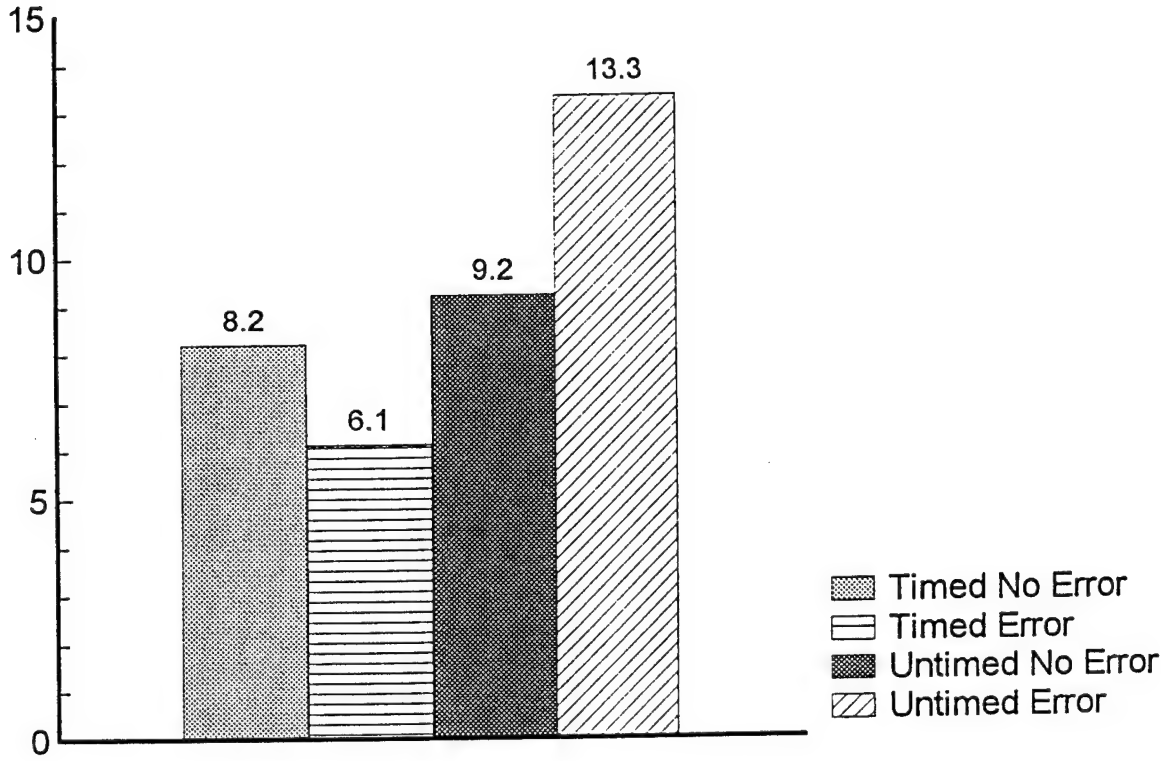


Figure 7. Experiment 5A--Timed vs. Untimed
Mean Number of Tests Conducted

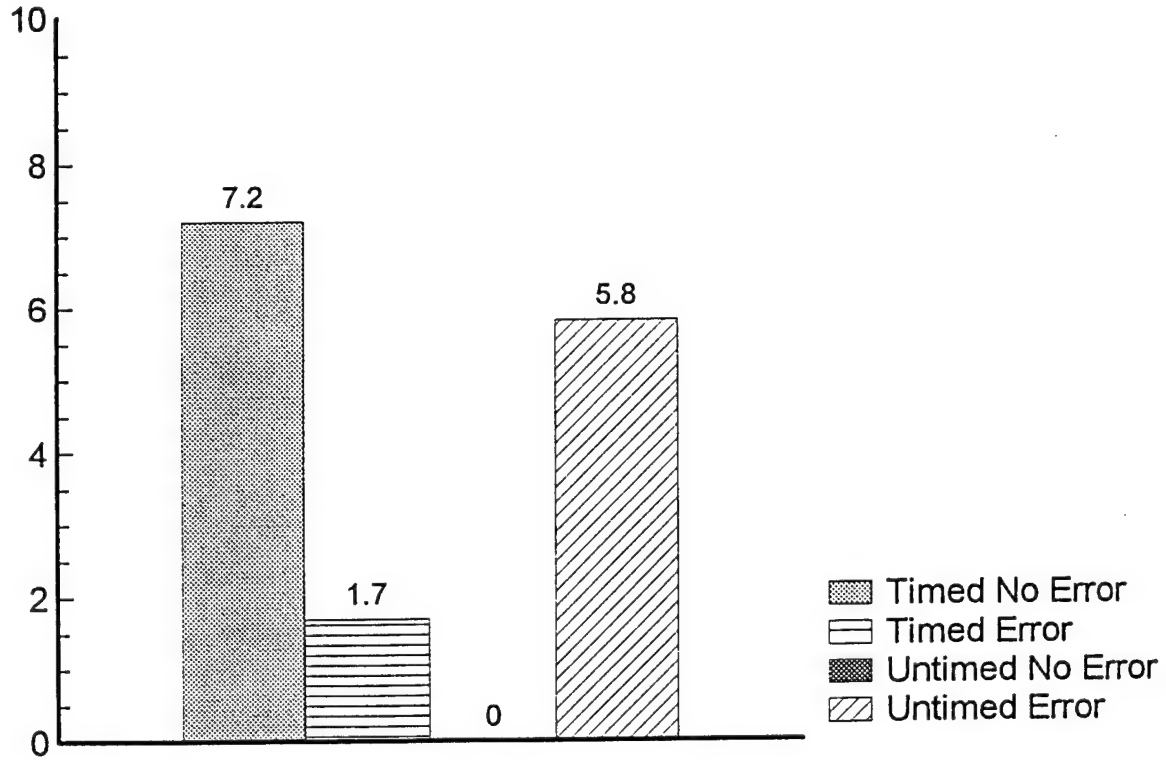


Figure 8. Experiment 5A--Timed vs. Untimed
Percentage of Expected Negative Test Outcomes

A two-way ANOVA indicated no significant main effects for time or error in the percentages of "unsure" responses ($F(1,31) = .011$, NS; $F(1,31) = 1.009$, NS). Subjects in the Timed No Error condition indicated that they were unsure of 16.7% of their total attempted test outcomes, while subjects in the Untimed No Error condition indicated that they were unsure of 22.8% of the outcomes. Subjects in the Timed Error condition were unsure of 10.0% of the outcomes, while subjects in the Untimed Error condition were unsure of 6.3% of the outcomes.

There were no significant main effects for time or error found among the mean percentages of confirmatory test outcomes ($F(1,31) = .008$, NS; $F(1,31) = .000$, NS). In the Timed No Error condition, 56.5% of the test outcomes were confirmatory, compared to 58.6% in the Untimed No Error condition. In the Timed Error condition, 59.6% of the test outcomes were confirmatory, compared to 55.6% in the Untimed Error condition. There were also no significant main effects for time or error found among the mean percentages of disconfirmatory test outcomes ($F(1,31) = .001$, NS; $F(1,31) = 2.448$, NS). In the Timed No Error condition, 26.9% of the test outcomes were disconfirmatory, compared to 18.6% in the Untimed No Error condition. In the Timed Error condition, 30.4% of the test outcomes were disconfirmatory, compared to 38.1% in the Untimed Error condition.

Discussion. The solving rates under the timed and untimed no error conditions (50.0% and 54.5% respectively) were very similar to the rates found for the 10 and 20-minute no error conditions in Experiment 3 (50.0% each). The imposed time limit (8 minutes) combined with data error had a marked detrimental effect on the solving rate (10.0%), as had been found in Experiment 3 in the 10 and 20-minute error conditions (30.0% and 20% respectively). Surprisingly, data error appeared to increase, rather than decrease, the rate of task solution for those subjects given unlimited time to discover the rule. The solving rate (75.0%) in the untimed error condition was considerably higher than the solution rate (40.0%) found in Walker and Harper's (1990) study and the solving rates found in Experiments 1 and 2 (55.6% and 46.6% respectively).

The mean number of tests (13.3) conducted by subjects in the untimed error condition was significantly higher than the mean numbers found for the other conditions and very similar to the mean number of tests (11.5) conducted by subjects in the 20-minute error condition in Experiment 3. Experimenters observed that subjects given unlimited time to solve the task usually spent about 20 minutes working on the problem--the same length of imposed time used in Experiment 3. The difference, of course, was in the use of a timer placed strategically in front of the subjects while they worked during Experiment 3. Given unlimited time to find the solution coupled with the error warning, subjects in Experiment 5 appeared to become more involved with

the task and better able to utilize the feedback information.

Though the number of subjects repeating tests across all conditions was fairly consistent, there was a decrease in the percentage of repeated tests to total tests (12.2%) used by subjects who did repeat tests in the untimed error condition. As had been discussed in Experiment 3, perhaps the information gleaned from repeated testing was detrimental rather than helpful in the error conditions.

Experiment 6A, Timed vs. Untimed; No Error vs. Error

Rationale. The Wason task was conducted under both timed and untimed normal and system failure conditions using the same methodology developed for Experiment 5. The purpose of the study was an attempt to replicate the earlier results using a more culturally diverse subject population.

Subjects. Sixty-six (66) University of Central Florida undergraduate students participated in the study. All subjects received experimental credit for their participation.

Procedure. Subjects were randomly assigned to one of four experimental conditions: (1) Untimed No Error; (2) Timed No Error; (3) Untimed Error; (4) Timed Error. The same procedure used in Experiment 5A was followed in administering the task.

Results. Nine out of 14 subjects (64.3%) in the Timed No Error condition and 12 out of 19 subjects (63.2%) in the Untimed No Error condition were able to solve the task. In the Timed Error condition, 12 out of 18 subjects (66.7%) were able to solve the task, compared to only 3 out of 15 subjects (20.0%) in the Untimed Error condition. The difference in solving rates among the conditions was significant ($\chi^2(3, N = 66) = 9.39, p = .025$). (See Figure 9.)

The mean number of tests conducted in the Timed No Error condition was 5.9, compared to 5.4 tests conducted in the Untimed No Error condition. The mean number of tests conducted in the Timed Error condition was 8.6, compared to 6.5 tests conducted in the Untimed Error condition. (See Figure 10.) A two-way ANOVA revealed a significant main effect for error in the number of tests conducted among the conditions ($F(1,62) = 7.964, p = .006$).

The mean percentage of total tests that were repeated in the Timed No Error condition was 47.4%, compared to 34.4% in the Untimed No Error condition. The mean percentage of total tests that were repeated in the Timed Error condition was 29.8%, compared to 20.6% in the Untimed Error condition. A two-way ANOVA revealed no significant main effects for time or error in the percentage of repeated tests conducted among the conditions ($F(1,62) = .41$, NS; $F(1,62) = .76$, NS). There was also no significant effect of condition on the number of individuals who repeated tests compared to those who did not repeat tests ($\chi^2(3, N = 66) = 6.29$, NS). Four out of 14 subjects (28.6%) in the Timed No Error condition repeated tests, compared to 5 out of 19 subjects (26.3%) in the Untimed No Error condition. In the Timed Error condition, 10 out of 18 subjects (55.5%) repeated tests, compared to 9 out of 15 Untimed Error condition subjects (60.0%).

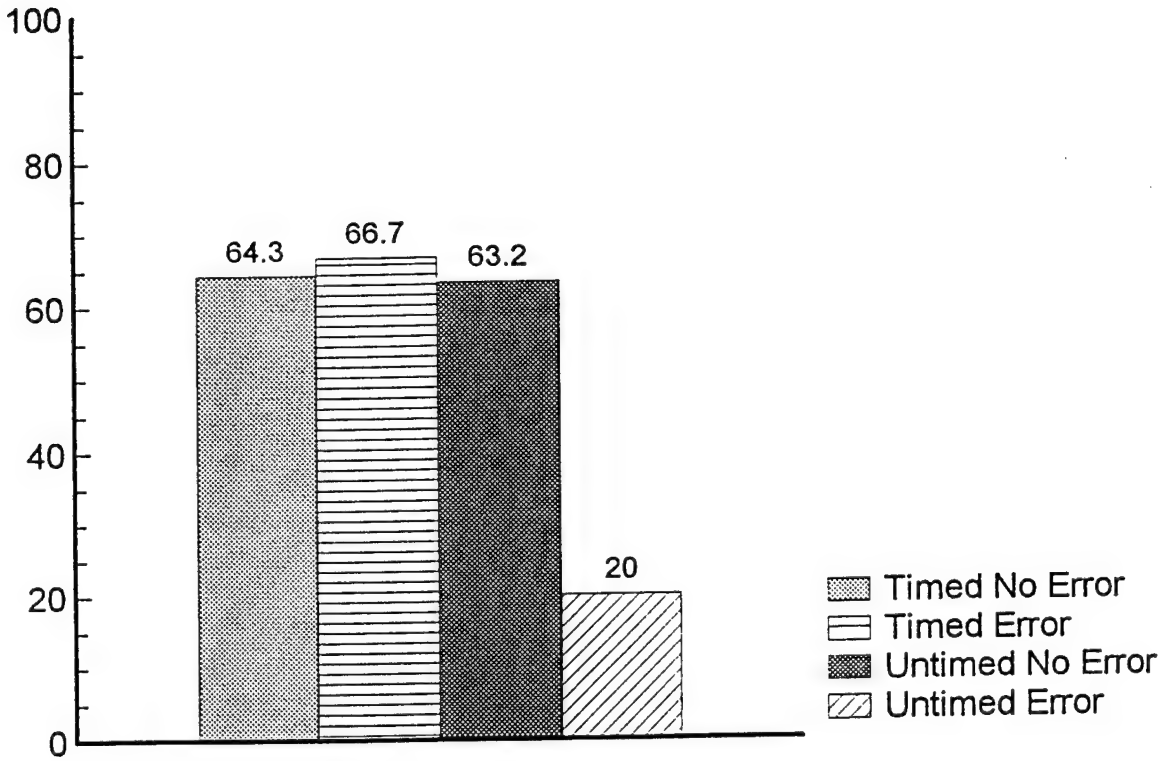


Figure 9. Experiment 6A--Timed vs. Untimed Percentage of Solvers

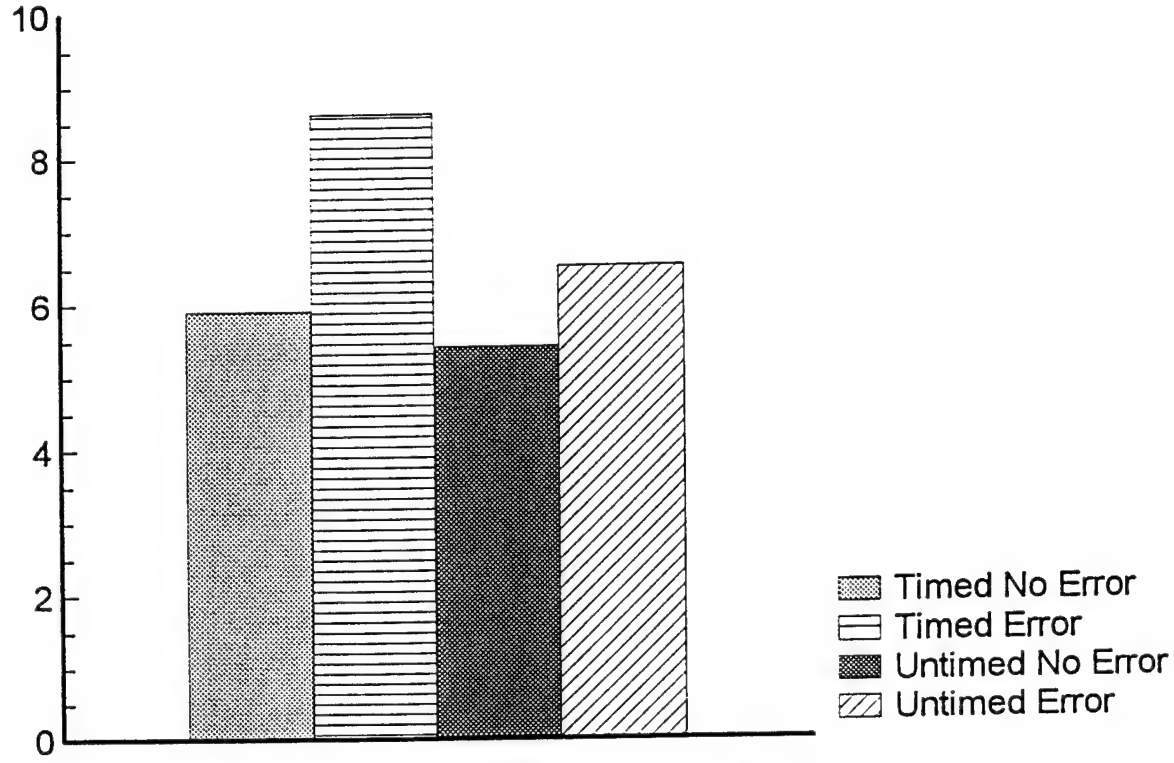


Figure 10. Experiment 6A--Timed vs. Untimed
Mean Number of Tests Conducted

A two-way ANOVA indicated no significant main effects for time or error in the percentages of total tests attempted that subjects' expected would fit the computer's rule ($F(1,62) = .440$, NS; $F(1,62) = .026$, NS). Subjects in the Timed No Error condition indicated that they expected 77.1% of their total attempted tests to fit the rule, while subjects in the Untimed No Error condition expected 82.7% to fit. Subjects in the Timed Error condition indicated that they expected 75.6% of their total attempted tests to fit the rule, while subjects in the Untimed Error condition expected 81.5% to fit.

A two-way ANOVA indicated no significant main effects for time or error among the percentages of total tests attempted that subjects' expected would not fit the computer's rule ($F(1,62) = 1.491$, NS; $F(1,62) = 1.010$, NS). Subjects in the Timed No Error condition indicated that they did not expect 1.4% of their total attempted tests to fit the rule, while subjects in the Untimed No Error condition did not expect 4.6% to fit. Subjects in the Timed Error condition indicated that they did not expect 4.1% of their total attempted tests to fit the rule, while subjects in the Untimed Error condition did not expect 8.1% to fit.

A two-way ANOVA indicated no significant main effects for time or error in the percentages of "unsure" responses ($F(1,62) = 1.344$, NS; $F(1,62) = .043$, NS). Subjects in the Timed No Error condition indicated that they were unsure of 21.4% of their total attempted test outcomes, while subjects in the Untimed No Error condition were unsure of 12.7% of the outcomes. Subjects in the Timed Error condition indicated that they were unsure of 20.3% of their total attempted test outcomes, while subjects in the Untimed Error condition were unsure of 10.4% of the outcomes.

There were no significant main effects for time or error found among the mean percentages of confirmatory test outcomes ($F(1,62) = 1.24$, NS; $F(1,62) = .498$, NS). In the Timed No Error condition, 57.1% of the test outcomes were confirmatory, compared to 61.9% in the Untimed No Error condition. In the Timed Error condition, 48.9% of the test outcomes were confirmatory, compared to 60.0% in the Untimed Error condition. There were also no significant main effects for time or error found among the mean percentages of disconfirmatory test outcomes ($F(1,62) = .066$, NS; $F(1,62) = 1.642$, NS). In the Timed No Error condition, 21.5% of the test outcomes were disconfirmatory, compared to 25.4% in the Untimed No Error condition. In the Timed Error condition, 30.8% of the test outcomes were disconfirmatory, compared to 29.6% in the Untimed Error condition.

Discussion. The percentages of solvers in the Timed and Untimed No Error conditions (64.3% and 63.2%) were only slightly higher than the percentages found in Experiment 5A (50.0% and 54.5%). The percentages of solvers in the Timed and Untimed Error conditions (66.7% and 20.0%), however, were a complete reversal

of the solving rates for the Timed and Untimed Error conditions (10.0% and 75.0%) found in Experiment 5A.

Other differences between the experimental results were also found. The mean numbers of tests conducted in the Timed and Untimed No Error conditions (5.9 and 5.4) were lower than the mean numbers found in Experiment 5A (8.2 and 9.2). The percentages of subjects repeating tests were much lower in both the Timed and Untimed No Error conditions (28.6% and 26.3%) than had been found in Experiment 5A (83.3% and 54.5%). However, the differences between the two experiments in the number of tests conducted and the use of repetition was not reflected by any major differences in solution rates for the no error conditions.

The mean numbers of tests conducted in the Timed and Untimed Error conditions (8.6 and 6.5) were lower than the mean numbers found in Experiment 5A (6.1 and 13.3), but the higher number of tests conducted in each of the experiments were consistent with the error conditions with higher solving rates. The percentages of subjects repeating tests were very similar in both the Timed and Untimed Error conditions (55.5% and 60.0%) to the percentages found in Experiment 5A (70.0% and 62.5%). Thus, it was felt in comparing the two experiments that successful solution under the error conditions might be related to the amount of data collected and that subjects enhanced performance under the Timed Error condition was again showing a greater degree of involvement in performing the task coupled with effective utilization of the test feedback.

Experiment 7A. Timed vs. Untimed; No Error vs. Error

Rationale. The Wason task was conducted under both timed and untimed normal and system failure conditions using the same methodology developed and used for Experiments 5A and 6A, except that subjects did not complete the sequence identification pretest. The purpose of the study was an attempt to replicate the results of Experiment 6A, in which system failure seriously disrupted task performance under an untimed condition and enhanced task performance under a timed condition.

Subjects. Seventy-six (76) University of Central Florida undergraduate students participated in the study. All subjects received experimental credit for their participation.

Procedure. Subjects were randomly assigned to one of four experimental conditions: (1) Untimed No Error; (2) Timed No Error; (3) Untimed Error; (4) Timed Error. The same procedures used in Experiments 5A and 6A was followed in administering the task, but four experimenters, rather than one, were used to run the subjects.

Results. In the Timed No Error condition, 13 out of 18 (72.7%%) subjects solved the task, compared to 14 out of 16 (87.5%) subjects in the Untimed No Error condition. In the Timed Error condition, 8 out of 23 (34.8%) subjects were able to solve the task, compared to 7 out of 19 (36.8%) in the Untimed Error condition. (See Figure 11.) The differences in solving rates among the conditions was significant ($\chi^2(3, N = 76) = 15.33$, $p = .002$).

The mean number of tests conducted in the Timed No Error condition was 6.2, compared to 5.6 in the Untimed No Error condition. The mean number of tests conducted in the Timed Error condition was 7.7, compared to 9.4 tests conducted in the Untimed Error condition. (See Figure 12.) A two-way ANOVA revealed a significant main effect for error in the number of tests conducted among the conditions ($F(1,72) = 7.584$, $p = .007$).

The mean percentage of total tests that were repeated in the Timed No Error condition was 21.3%, compared to 26.7% in the Untimed No Error condition. The mean percentage of total tests that were repeated in the Timed Error condition was 27.0%, compared to 28.1% in the Untimed Error condition. A two-way ANOVA indicated no significant main effects for time or error in the percentage of repeated tests conducted among the conditions ($F(1,72) = .338$, NS; $F(1,72) = 2.623$, NS). However, there was a significant effect of condition on the number of individuals who repeated tests compared to those who did not repeat tests ($\chi^2(3, N= 76) = 8.719$, $p = .03$). In the Timed No Error condition, 5 out of 18 subjects (27.8%%) repeated tests, compared to 4 out of 16 subjects (25.0%) in the Untimed No Error

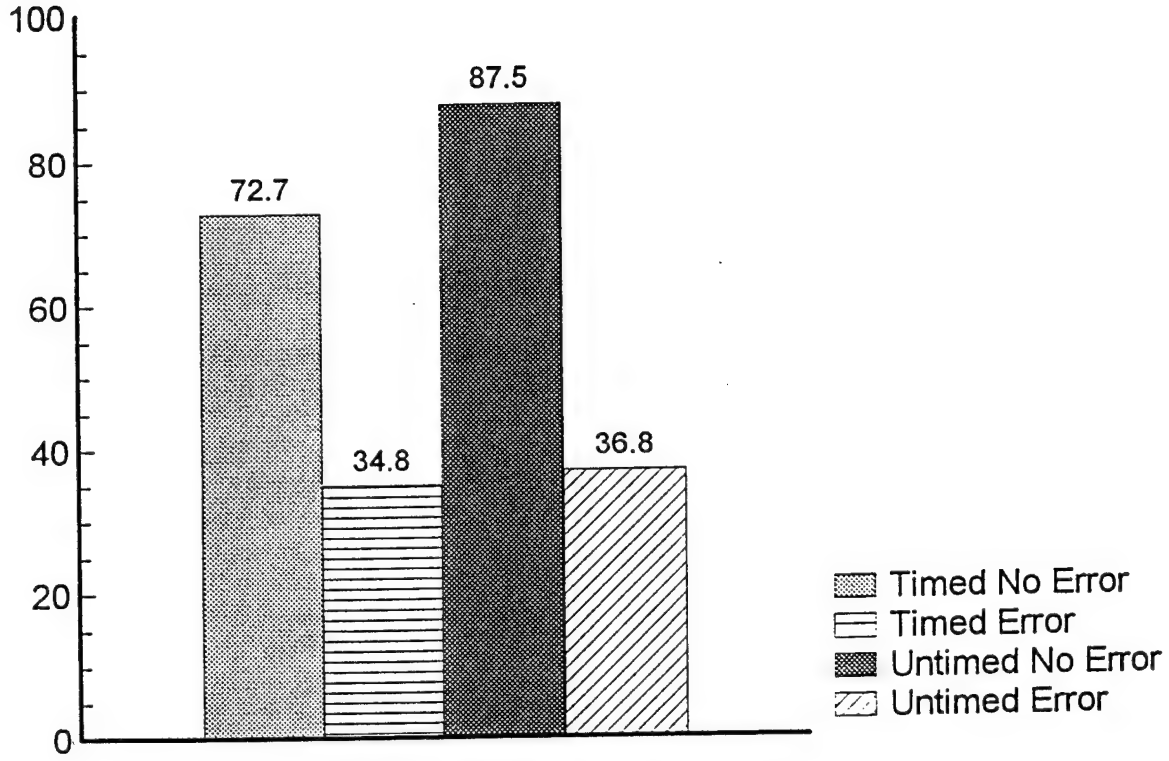


Figure 11. Experiment 7A--Timed vs. Untimed Percentage of Solvers

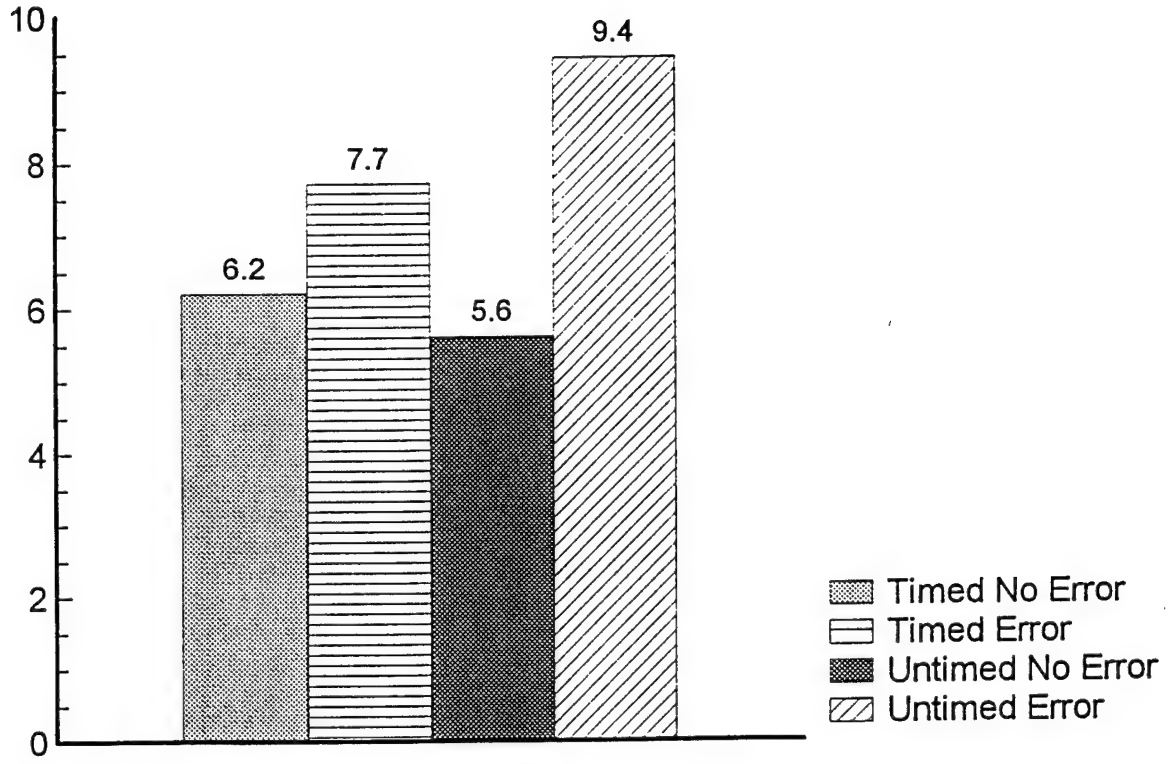


Figure 12. Experiment 7A--Timed vs. Untimed
Mean Number of Tests Conducted

condition. In the Timed Error condition, 15 out of 23 subjects (65.2%) repeated tests, compared to 7 out of 19 (36.8%) in Untimed Error condition. (See Figure 13).

A two-way ANOVA indicated no significant main effects for time or error in the percentages of total tests attempted that subjects' expected would fit the computer's rule ($F(1,72) = .009$, NS; $F(1,72) = .417$, NS). Subjects in the Timed No Error condition indicated that they expected 73.5% of their total attempted tests to fit the rule, while subjects in the Untimed No Error condition indicated that they expected 72.6% of their total attempted tests to fit. In the Timed Error condition subjects expected 78.5% to fit the rule, while subjects in the Untimed Error condition expected 77.9% of their total attempted tests to fit.

A two-way ANOVA also indicated no significant main effects for time or error in the percentages of total tests attempted that subjects' did not expect would fit the computer's rule ($F(1,72) = .549$, NS; $F(1,72) = .945$, NS). Subjects in the Timed No Error condition did not expect 2.9% of their total attempted tests to fit the rule, while subjects in the Untimed No Error condition did not expect 2.1% of their total attempted tests to fit the rule. Subjects in the Timed Error condition did not expect 2.6% to fit, while subjects in the Untimed Error condition did not expect 6.7% to fit.

A two-way ANOVA indicated no significant main effects for time or error in the percentages of "unsure" responses ($F(1,72) = .013$, NS; $F(1,72) = .925$, NS). Subjects in the Timed No Error condition indicated that they were unsure of 23.6% of their total attempted test outcomes, while subjects the Untimed No Error condition indicated that they were unsure of 25.3% of their total attempted test outcomes. Subjects in the Timed Error condition indicated they were unsure of 18.9% of the outcomes, while subjects in the Untimed Error condition indicated they were unsure of 15.4% of the outcomes.

There were no significant main effects for time or error found among the mean percentages of confirmatory test outcomes ($F(1,72) = .087$, NS; $F(1,31) = .533$, NS). In the Timed No Error condition, 55.9% of the test outcomes were confirmatory, compared to 54.1% of the test outcomes in the Untimed No Error condition. In the Timed Error condition, 47.6% of the test outcomes were confirmatory, compared to 53.1% in the Untimed Error condition. A significant main effect for error was found among the mean percentages of disconfirmatory test outcomes ($F(1,72) = 5.706$, $p = .02$). (See Figure 14.) In both the Timed and Untimed No Error condition, 20.6% of the test outcomes were disconfirmatory. In the Timed Error condition, 33.5% of the test outcomes were disconfirmatory, compared to 31.5% in the Untimed Error condition.

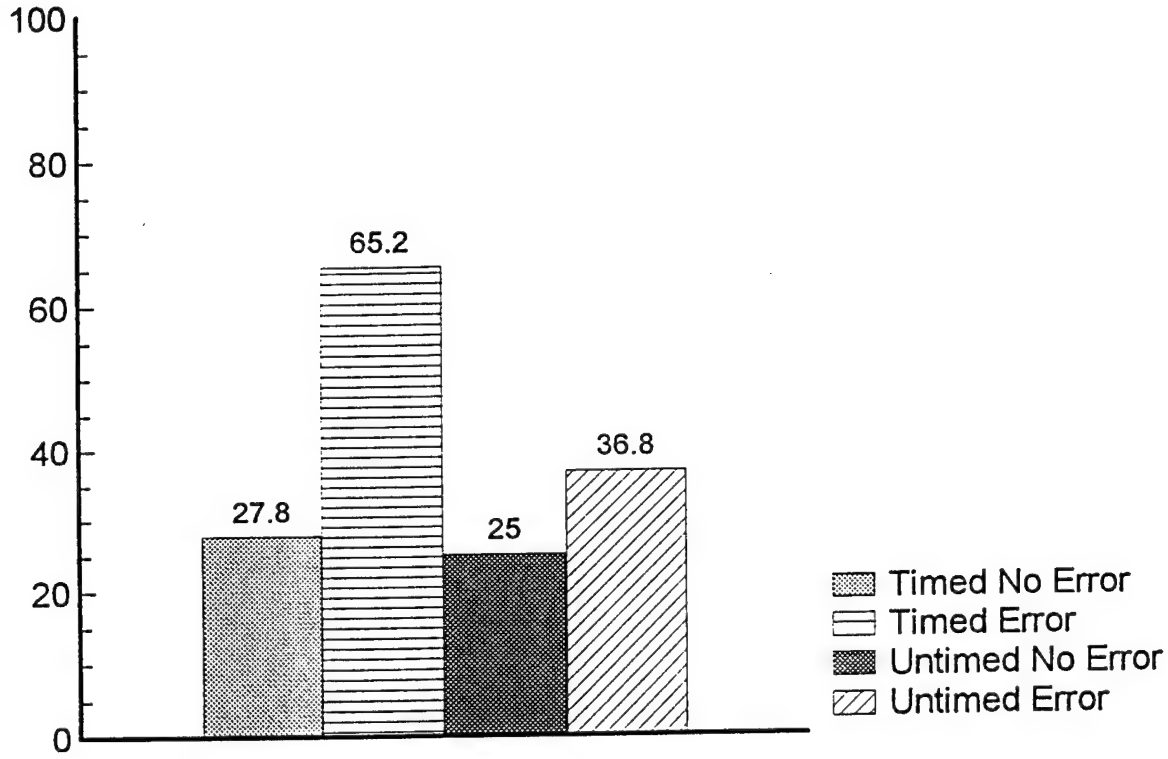


Figure 13. Experiment 7A--Timed vs. Untimed
Percentage of Subjects Repeating Tests

Discussion. Solving rates for both the Timed and Untimed No Error conditions (72.2% and 87.5%) were higher than had been found in Experiments 5A and 6A. In contrast, the solving rate for the Timed Error condition (34.8) was much lower than had been found in Experiment 6A (66.7%) and slightly higher than had been found in Experiment 5A (10.0%). The solving rate for the Untimed Error condition (36.8%) was slightly higher than had been found in Experiment 6A (20.0%) and much lower than had been found in Experiment 5A (75.0%). (See Table 1 for complete performance comparisons across all three experiments.) Due to the similarities in the solving rates for both the Timed and Untimed Error conditions, data error, and not time, appeared to contribute to decreased task performance.

As had been found in both Experiments 5A and 6A, subjects in the Error conditions conducted significantly more tests than those in the No Error conditions. Unlike Experiments 5A and 6A, conducting a higher number of tests was not consistent with an increase in solving rates. (See Table 2 for complete heuristic comparisons across all three experiments.) One possible explanation for the decreased solving rate in the Untimed Error condition despite the increased number of overall tests concerns the number of subjects repeating tests. In the Untimed Error condition in Experiment 5A, a greater percentage of subjects repeated tests and also used a higher number of tests. In the current experiment, a much lower percentage of subjects repeated tests (36.8%) and also used a higher number of tests. Therefore, subjects in the Untimed Error condition were not checking for error by test repetition and were receiving much more potentially confusing information.

Phase IIA. General Discussion. The comparison of solving rates in the No Error conditions across all five experiments conducted in Phase IIA demonstrated fairly consistent performance in both the Timed and Untimed conditions. Such performance rates were also consistent with earlier studies under untimed conditions (Walker & Harper, 1989; Walker, 1987). Furthermore, consistency was also demonstrated in the mean number of tests conducted, mean percentages of repeated tests, and mean percentages of consistent tests. Thus, it appeared that the solution rates and problem-solving behaviors remained stable for the majority of subjects tested under both untimed and timed no error conditions.

The introduction of an imposed time limit in Experiments 3A, 5A, and 7A had significant and fairly consistent detrimental effects on problem-solving performance. The enhanced performance under system failure and time limitations found in Experiment 6A was not replicated in Experiment 7A. It was felt that the finding should be considered an experimental anomaly. In all four experiments, consistency was again demonstrated among the timed and untimed error conditions in the mean number of tests conducted, mean percentages of repeated tests conducted, and mean

percentages of consistent tests conducted. In addition, the problem-solving behaviors were also similar to those found in the timed and untimed no error conditions. Thus, the introduction of system failure and time limitations did not appear to change how subjects approached the task. The failure to alter problem-solving behavior in response to data error might, therefore, have resulted in the decrease in solving rates.

Table 1

Performance Comparisons

No Error Conditions

	Untimed			Timed		
Sample	CSU	UCF		CSU	UCF	
Experiment	5A	6A	7A	5A	6A	7A
	(N=11)	(N=19)	(N=16)	(N=6)	(N=14)	(N=18)
Percent Solved	54.5%	63.2%	87.5%	50.0%	64.3%	72.2%

Error Conditions

	Untimed			Timed		
Sample	CSU	UCF		CSU	UCF	
Experiment	5A	6A	7A	5A	6A	7A
	(N=8)	(N=15)	(N=19)	(N=10)	(N=18)	(N=23)
Percent Solved	75.0%	20.0%	36.8%	10.0%	66.7%	34.8%

Table 2

Heuristics Comparisons

No Error Conditions

	Untimed			Timed		
Sample	CSU	UCF		CSU	UCF	
Experiment	5A	6A	7A	5A	6A	7A
Sample Size	(N=11)	(N=19)	(N=16)	(N=6)	(N=14)	(N=18)
Mean No. of Tests Conducted	9.2	5.4	5.6	8.2	5.9	6.2
% of Subjects Repeating Tests	54.5%	26.3%	25.0%	83.3%	28.6%	27.8%
Mean % of Repeated Tests	37.9%	34.4%	26.7%	21.8%	47.4%	21.3%
Expected Outcomes						
Mean % of "Yes"	77.2%	82.7%	72.6%	76.1%	77.2%	73.5%
Mean % of "No"	0.0%	4.6%	2.1%	7.2%	1.4%	2.9%
Mean % of "Unsure"	22.8%	12.7%	25.3%	16.7%	21.4%	23.6%

Error Conditions

	Untimed			Timed		
Sample	CSU	UCF		CSU	UCF	
Experiment	5A	6A	7A	5A	6A	7A
Sample Size	(N=8)	(N=15)	(N=19)	(N=10)	(N=18)	(N=23)
Mean No. of Tests Conducted	13.3	6.5	9.4	6.1	8.6	7.7
% of Subjects Repeating Tests	62.5%	60.0%	36.8%	70.0%	55.5%	65.2%
Mean % of Repeated Tests	12.2%	20.6%	28.1%	28.6%	29.8%	27.0%
Expected Outcomes						
Mean % of "Yes"	88.0%	81.5%	77.9%	88.3%	75.6%	78.5%
Mean % of "No"	5.8%	8.1%	6.7%	1.7%	4.1%	2.6%
Mean % of "Unsure"	6.3%	10.4%	15.4%	10.0%	20.3%	18.9%

Phase IIB.--Artificial Universe Studies (Tribbles Task)

Experiment 5B. No Error vs. Error

Rationale. Part B of the fifth experiment was designed to assess subjects' problem-solving behavior using an artificial universe task, "Tribbles". In this task, subjects were asked to pilot a spaceship above the surface of a planet and drop imaginary life forms, "Tribbles", from the ship to determine which part of the planet could support life. Feedback concerning whether or not the "Tribbles" lived or died was provided after each drop.

Subjects. Forty-two (42) Central State University subjects participated in the experiment. All of these subjects had also participated in Experiment 5A.

Procedure. All subjects were given a sheet of instructions which explained that they were being asked to discover a moisture boundary line on an imaginary planet's surface. The moisture boundary line determined the resulting life or death of life forms dropped from the subject's spaceship. All subjects were given twelve "Tribbles" to drop from their spaceship. After each drop, feedback concerning the success of the drop was given. The subject was able to move a hypothesized boundary line anytime during the task. Following the last drop, the subject was asked to position the line where they felt it should be, based on the results of the feedback.

For some conditions, subjects were also given system failure (false feedback), measurement error (indefinite drop locations), or a combination of the two. In both the system failure and measurement error conditions, subjects were given a limited number of probes to use to check the drop data.

Results. Three out of four subjects (75.0%) in the normal condition solved the task. Four out of eight subjects (50.0%) given only system failure solved the task compared to three out of twelve subjects (25.0%) given measurement error. Four out of eighteen subjects (22.2%) given both system failure and measurement error solved the task. (See Figure 14.) However, the differences in solving rates among the conditions was not significant.

Conclusion. Though the differences among the conditions was not statistically significant, it was apparent that both measurement error and the combination of measurement error and system failure had a detrimental effect on subjects' performance.

Discussion. It was not clear whether or not subjects' clearly understood the concepts of system failure and measurement error introduced in the task. Experiment 6B was designed to look only at the effects of system failure.

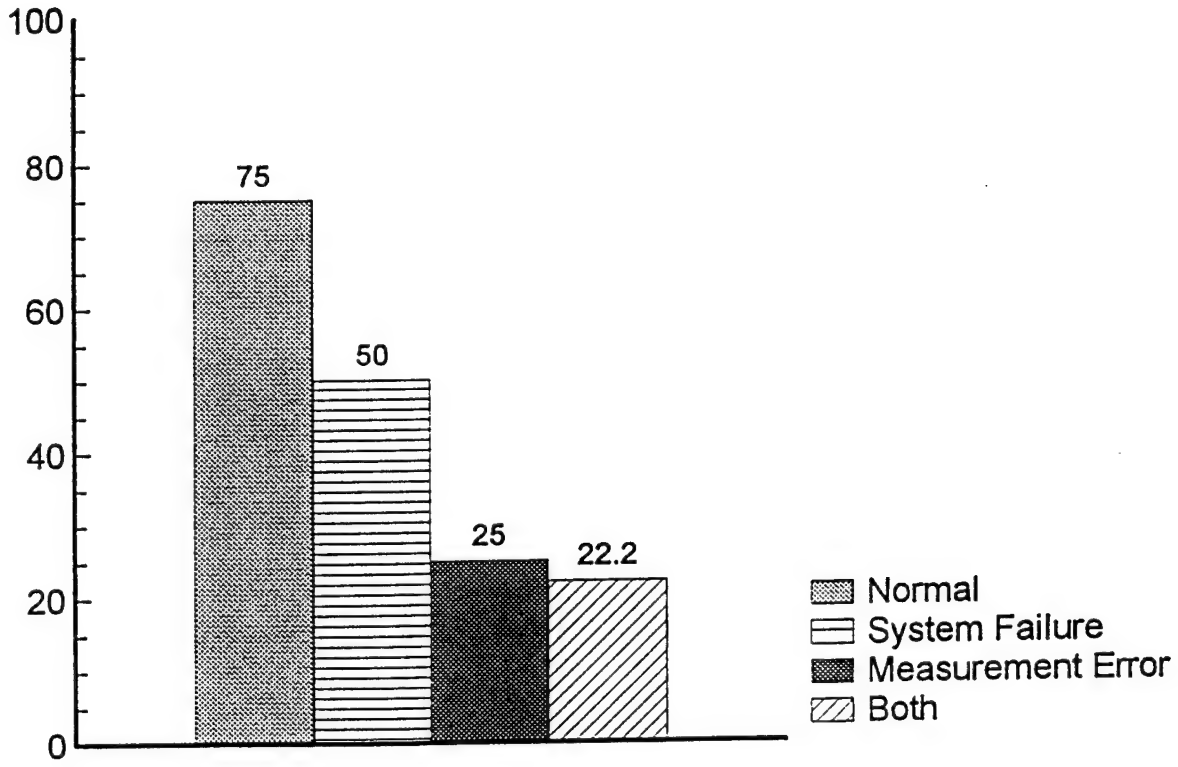


Figure 14. Experiment 5B--Tribbles, No Error vs. Error Percentage of Solvers

Experiment 6B. No Error

All 73 University of Central Florida students who participated in Experiment 6A also completed the Tribbles task. Due to several errors found in the original programming for the system failure condition, the task was only administered under the normal condition. Out of the 73 subjects, 61 (83.6%) were able to accurately locate the hypothetical moisture line.

Experiment 7B. No Error vs. High and Low Feedback Error

Rationale. The purpose of the experiment was to assess the effect of low and high system failure on subjects' ability to use a limited amount of data to correctly determine a parameter.

Subjects. Seventy-three (73) undergraduate students participated in the study. All subjects received experimental credit for their participation.

Procedure. Subjects were randomly assigned to one of three possible system failure conditions--no system failure, low system failure, and high system failure. All subjects were asked to discover a moisture boundary line on an imaginary planet. To discover the line, they were given 12 moisture-dependent creatures to drop on the planet's surface. After each drop, they received feedback as to whether the creature lived or died. In the low system failure condition, the feedback from two of the twelve tests was reversed. In the high system failure condition, the feedback from three of the twelve tests was reversed. For both system failure conditions, the subjects were informed that the feedback might be wrong and they were given two probes to check questionable data. Following the last test, the subjects were required to locate the boundary line.

Results. Nine (9) out of 12 subjects (75.0%) in the no system failure condition discovered the location of the boundary line, compared to 13 out of 40 subjects (32.5%) in the low system failure condition and 4 out of 21 subjects (19.0%) in the high system failure condition. The difference in solving rates among the conditions was significant ($\chi^2 = 10.8$, df = 2, p = .005). (See Figure 15.)

A significant difference was also found between the low and high system failure conditions in the degree of pixel difference between the correct line and subjects' incorrect solutions. The mean pixel distance from the correct line in the low system failure condition was 72.33, compared to 174.18 in the high system failure condition ($t(42) = 2.951$, $p = .005$).

Discussion. While system failure seriously disrupted task performance in both error conditions, it should be noted that

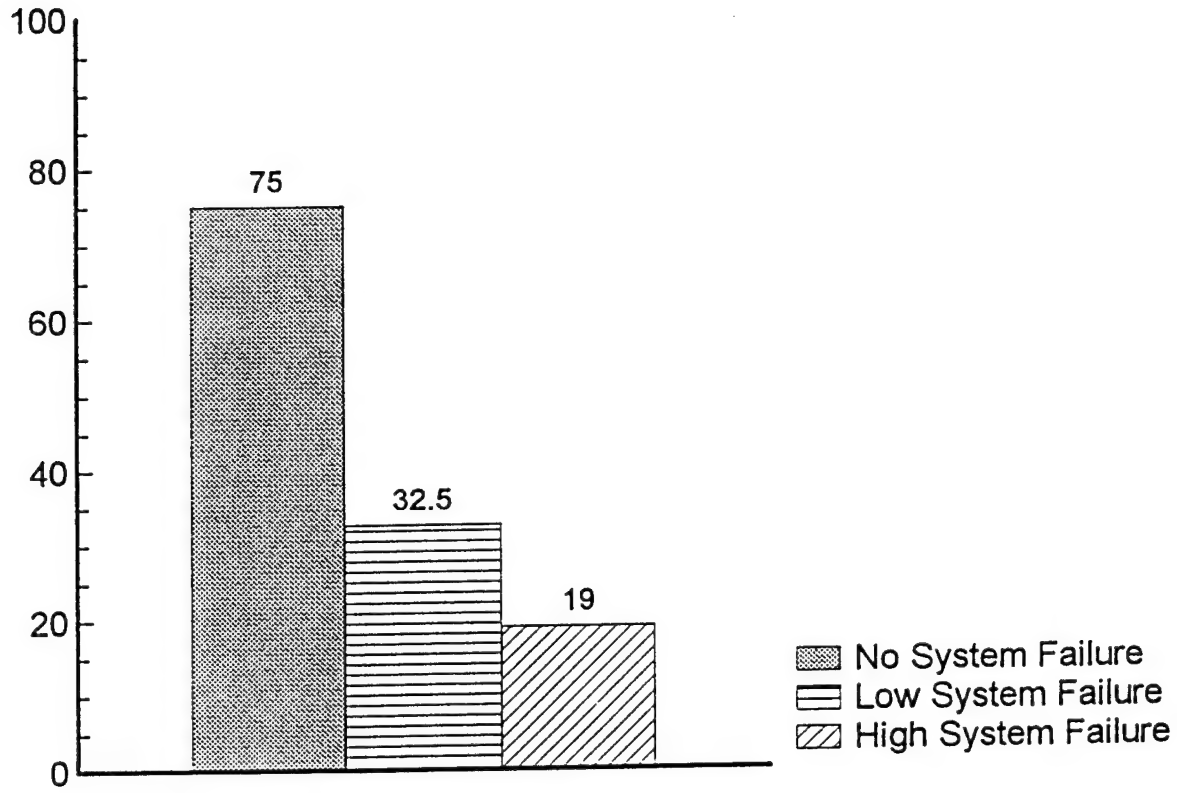


Figure 15. Experiment 7B--Tribbles, Percentage of Solvers

increasing the level of system failure by merely one additional error trial (from low to high) significantly decreased subjects' ability to narrow down the correct parameter.

Experiment 8. Protocol Analysis--No Error vs. Error

Rationale. The purpose of the experiment was to use protocol analysis to discover whether or not subjects were thinking about the possibility of error and what problem-solving strategies were being utilized.

Subjects. Twenty-two (22) University of Central Florida undergraduate students were recruited for the study. All were given experimental credit for their participation.

Procedure. The experiment utilized the same basic procedures for the Wason 2-4-6 task under Untimed No Error and Error conditions that were followed in Experiments 5A, 6A, and 7A. However, before beginning the actual task, all subjects were given a warm-up task during which they were asked to "think aloud" (Ericsson & Simon, 1984) while solving the multiplication problem, "24 x 34". Subjects were then randomly assigned to one of two experimental conditions: (1) No Error or (2) Error and required to complete a minimum of ten trials. All subjects were tape-recorded as they verbalized their thoughts while solving the task.

Transcripts of each subject's tape-recorded verbalizations were segmented into units signifying a complete problem-solving process, such as "hypothesis generation" or "strategy formation." Each unit was then segmented into the distinct propositions it incorporated. The purpose of the segmentation was to divide up the protocols "...so that each segment will constitute one instance of a general process" (Ericsson and Simon, 1984, p. 205). For instance, the two statements--"I think I'll try 'even numbers'. I'll test 8-10-12."--would be considered one complete unit. The complete unit was then divided into two segments corresponding to the two distinct propositions--1(A) "I think I'll try..." and 1(B) "I'll test...".

The segmented transcripts and computer printouts were integrated and encoded using a list of operators (see Figure 16), which was modified from a list developed for a previous protocol analysis study of the Wason task (Walker, 1985). An operator was considered to be "...a process that generates or transforms knowledge" (Ericsson & Simon, 1984, p. 175). Walker (1985) stated that "...the basic solution process involved proposing and testing hypotheses, which, on the basis of positive or negative experimenter feedback, eventually led to a rule announcement" (p. 7). In the current study, the basic solution process remained the same except that subjects selected, rather than proposed hypotheses, and based their eventual rule announcement on positive or negative computer feedback as opposed to experimenter feedback. Thus, the segment (1A) "I think I'll try 'even numbers'" would be coded as "HYP" (hypothesis selection) and the segment, (1B) "I'll test 8-10-12", would be coded as "CTEST"

HYP - Selected and/or stated hypothesis about rule.

CTEST - Consistent test (matches hypothesis).

ITEST - Inconsistent test (does not match hypothesis).

NTEST - Test without stated hypothesis.

PRED - Predicted test outcome.

RESULT - Actual test outcome.

READ - Reading screen instructions.

SCLAR - Subject query for task clarification.

ECLAR - Experimenter response to subject query.

PROMPT - Experimenter prompt to "think-aloud".

TCOMM - Task comment.

RCOMM - Result comment.

NCOMM - Non-task comment.

STRAT - Stated strategy for solving task.

REV - Review of previous tests and results.

GUESS - Final rule guess.

Figure 16. State coding operators

(consistent number-sequence test). The test was considered consistent since the test sequence matched the selected hypothesis. (See Figure 17.)

Using a problem behavior graph key (see Figure 18), encoded transcripts were converted into graphs (see Figures 19 through 22) for comparison. Each graph indicated the direction and continuity of the subject's problem-solving process by tracing the flow of hypotheses selected and the test sequences used to investigate them. In keeping with Ericsson and Simon (1984), "...the analysis assumes that the subject solves the problem by searching through one or more problem spaces (i.e., sets of alternative states of knowledge" (p. 195). Details about each step in the decision-making process (type of test, number sequence used) were included in the graphs. Horizontal tracking of the hypotheses selected and tests conducted was used when subjects sequentially selected, but did not rule out competing hypotheses. Vertical tracking from a previously selected hypothesis was indicative of hypothesis elimination. If a subject returned to a previously eliminated hypothesis or repeated a test, the corresponding geometric figure was shaded. The graph was considered complete when the subject gave a final statement of the hypothesis.

General Results. In the No Error condition, 9 out of 11 (81.8%) subjects were able to solve the task, compared to 5 out of 11 (45.5%) subjects in the Error condition (see Figure 23). The difference in solving rate proportions between the conditions was significant ($z = 2.70$, $p < .01$, two-tailed).

The mean number of tests conducted by subjects in the No Error condition was 9.9, compared to 17.3 trials for the Error condition. The difference in the mean number of total attempted tests between the conditions was significant ($t(20) = 2.292$, $p = .03$).

The mean percentage of total tests that were repeated sequences in the No Error condition was 29.9% compared to 21.8% in the Error condition. The difference in the mean percentages of repeated tests between the conditions was not significant ($t(20) = .123$, NS). There was also no significant difference between conditions in the proportions of subjects who repeated tests ($z = -.676$, NS, two-tailed). In the No Error condition, 7 out of 11 (63.6%) subjects repeated tests compared to 9 out of 11 (81.8%) subjects in the Error condition.

There was no significant difference between conditions in the mean percentages of total tests attempted that subjects' expected would fit the computer's rule ($t(20) = .178$, NS). Subjects in the No Error condition expected 63.7% of their attempted tests to fit the rule, while subjects in the Error condition expected 66.3% to fit. The difference between conditions in the mean

Subject 408 (Error, Not Solved)

TCOMM - ALRIGHT MY #'S ARE 2-4-6 AND THEY'RE EVEN #'S...
(HYP - EVEN #'S)
CTEST - LET'S SEE THE SEQUENCE, 4, 6, AND 8...AND YES IT IS...
PRED - AND YES I DO YES...
RESULT - AND THE COMPUTER SAYS YES, OK...UM...UM...
HYP - I THINK I'LL DO EVEN NUMBERS AGAIN...
CTEST - AND I'LL DO 12, 14, AND 16...YES,
PRED - AND UM UNSURE...
RESULT - IT SAYS NO...UM...OK,
HYP - SO ... I'M GONNA GO NUMBERS LESS THAN 10...
CTEST - AND I'M GOING TO DO 3, 5,7 OK...AND YES,
PRED - AND NO I DON'T THINK IT FITS,
RESULT - IT SAID NO (ERROR TRIAL)
HYP - OK, I'M GOING TO DO...UM, I'M GOING TO DO ASCENDING #'s
CTEST - UM I'M GOING TO SAY 2,6,AND 8...
TCOMM - AND THERE ARE ANY (?) AND THEY'RE NOT IN ORDER, SO, YES,
PRED - AND UM..I'M GOING TO SAY YES
(RESULT- YES)
TCOMM - UM...OK...OTHER...
HYP - I'M GOING TO SAY ASCENDING EVEN NUMBERS
CTEST - AND I'M GONNA GO... 2,4,8 AS MY SEQUENCE, YES
PRED - AND YES,
(RESULT- YES)
HYP - OK...I'M GOING TO TRY NUMBERS LESS THAN 10 (REPEAT)
STRAT - AND JUST SEE IF ODD REALLY DIDN'T WORK,
CTEST - SO I'M GONNA GO 3,4, AND 5, AND SEE IF IT TAKES
PRED - UNSURE,
RESULT - IT SAYS IT FITS,

Figure 17. Integrated transcript and computer print-out example

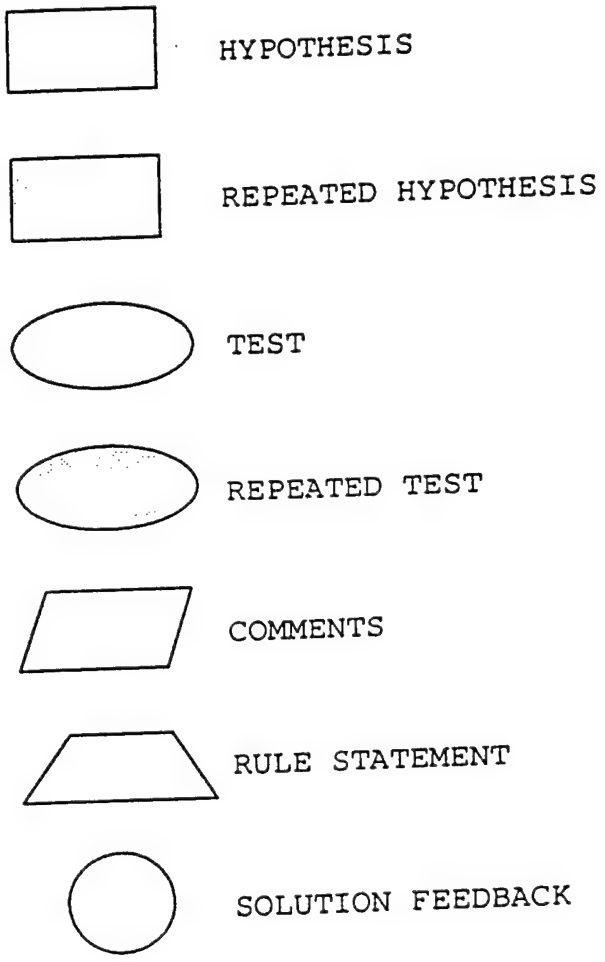


Figure 18. Problem behavior graph key

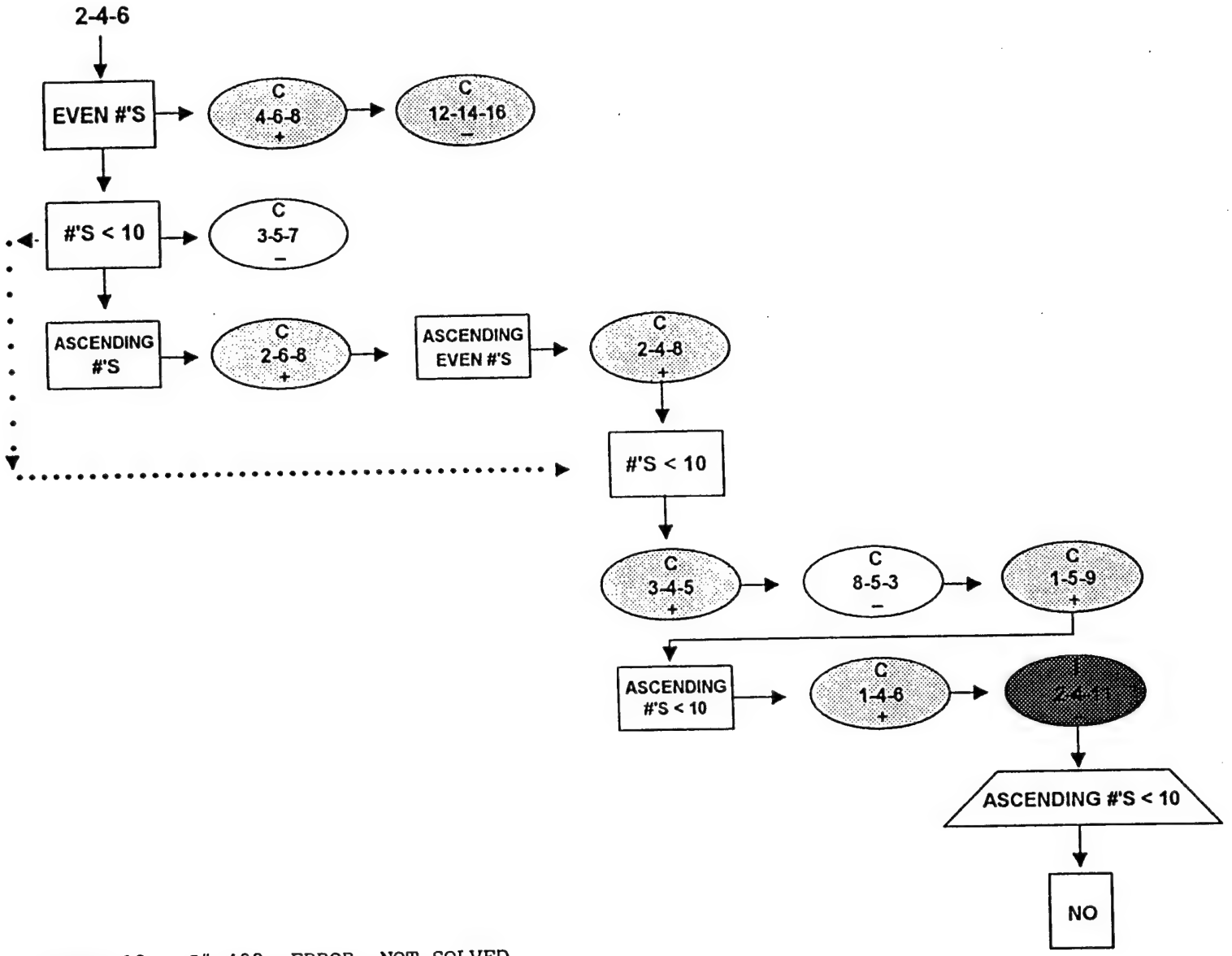
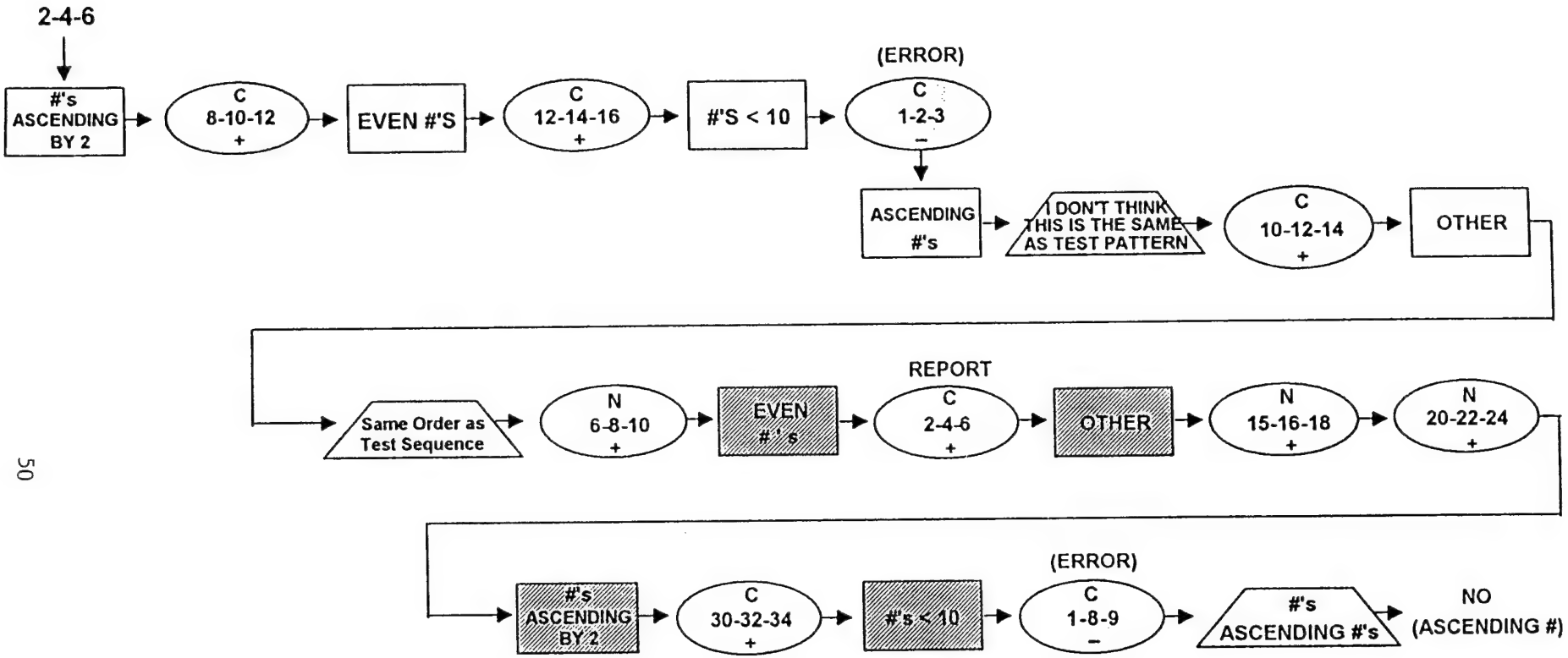
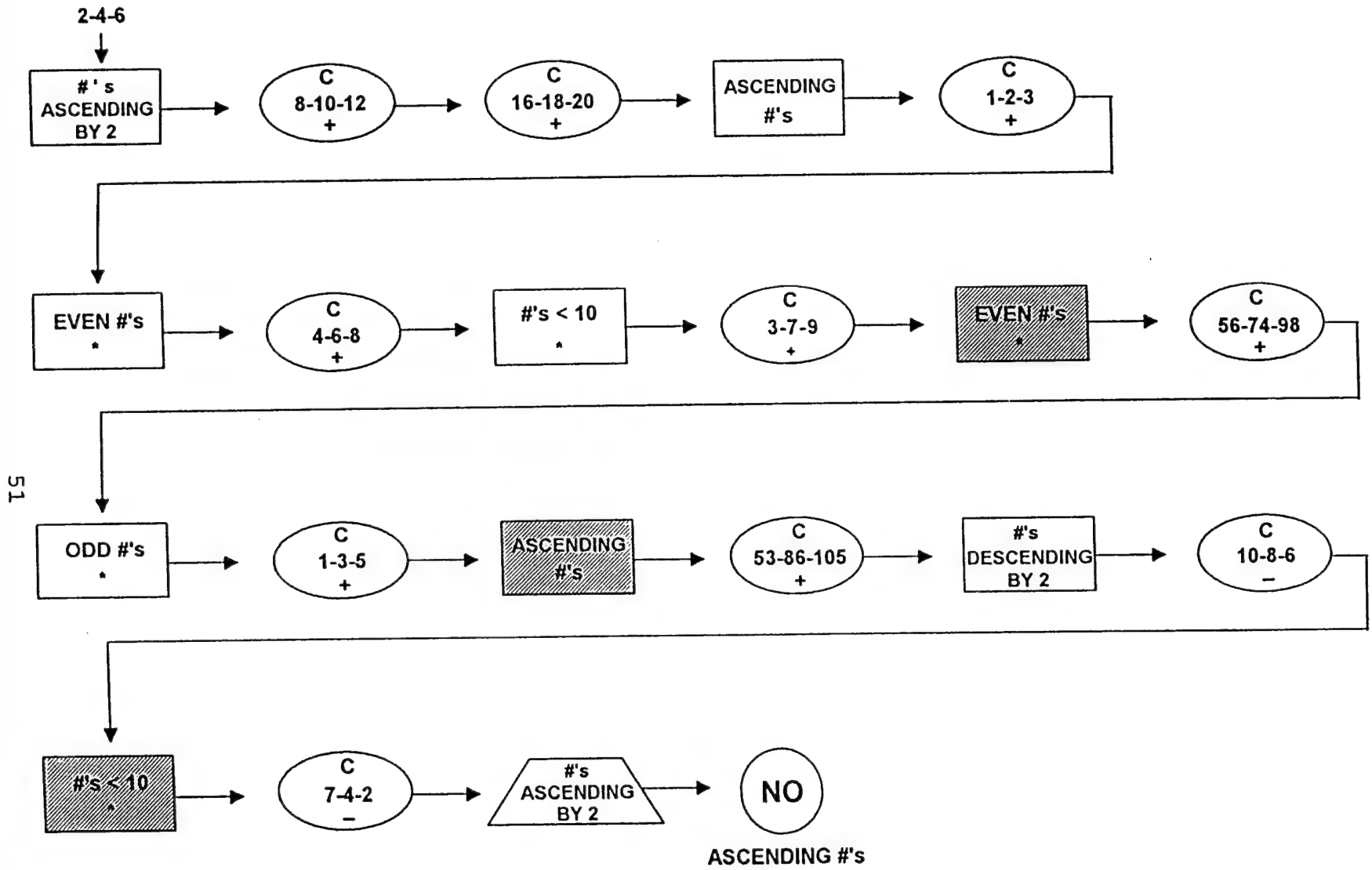


FIGURE 19. S# 408, ERROR, NOT SOLVED



*NOTE: NO REFERENCE IN PROTOCOL TO POSSIBILITY OF ERROR

FIGURE #20. PBG, S#419, ERROR, NOT SOLVED



*NOTE: CHOSE A RULE THAT WAS ALREADY ELIMINATED BY A TEST.

FIGURE 21. PBG, S# 402, NO ERROR, NOT SOLVED

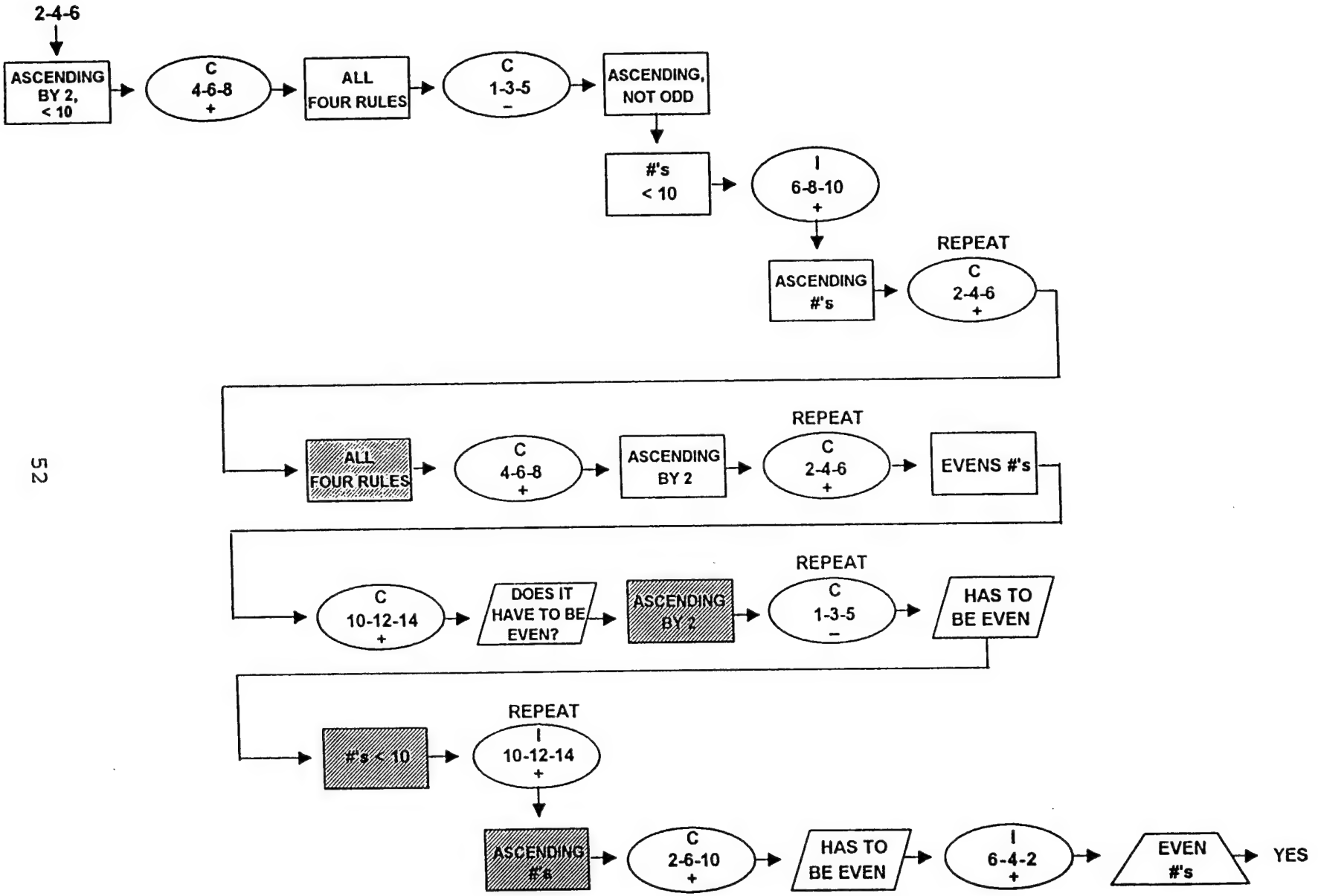


FIGURE 22. PBG, S# 420, NO ERROR, SOLVED

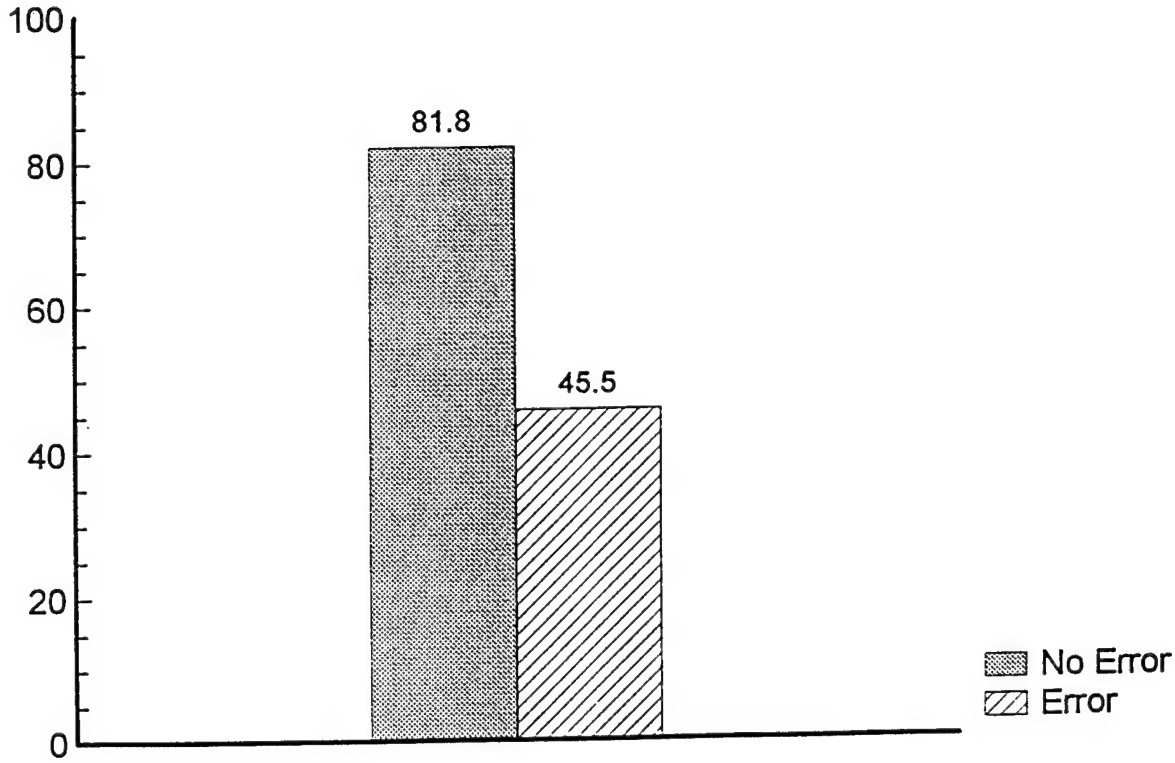


Figure 23. Experiment 8--Protocol Analysis, No Error vs. Error Percentage of Solvers

percentages of total tests attempted that subjects did not expect to fit the computer's rule was also not significant ($t(20) = .146$, NS). Subjects in the No Error condition expected 15.5% of their total attempted tests to fit the rule, while subjects in the Error condition expected 14.2% not to fit. Similarly, the difference between conditions in the mean percentages of "Unsure"responses was not significant ($t(20) = .095$, NS). Subjects in the No Error condition indicated they were unsure of 20.8% of the attempted test outcomes while subjects in the Error condition were unsure of 19.5% of the test outcomes.

No significant difference was found between the mean percentages of confirmatory test outcomes in the No Error condition (65.3%) and the Error condition (47.3%) ($t(20) = .095$, NS). However, a significant difference was found between the mean percentages of disconfirmatory test outcomes in the No Error condition (13.9%) and the Error condition (33.2%) ($t(20) = 2.732$, $p = .013$).

Problem Behavior Graph Analyses. Overall analyses of the graphs indicated that most subjects entertained several hypotheses at once by sequentially selecting potential rules and conducting tests which matched the selections. For example, Subject 409 sequentially tested the hypotheses "Numbers less than 10", "Numbers ascending by 2", "Even numbers", and "Ascending numbers" with matching tests. Negative feedback from the test "1-2-3" was then used to eliminate "Ascending numbers". The majority of subjects also tested all stated rules available on the selection list.

While hypotheses were tested primarily using tests which matched the selection, many such tests could be used to eliminate other potential rules. Six out of nine solvers in the No Error condition expressed an understanding of how negative feedback to a number sequence test eliminated other rules. For instance, following a review of previously selected rules and test results, Subject 409 stated that "ascending numbers had been eliminated". In addition, all No Error solvers appeared to utilize negative feedback, rechecked rules with different tests, and some also repeated previous tests. In contrast, No Error non-solvers clearly ignored data which disconfirmed final rule selection. For example, Subject 402 declared "Numbers ascending by 2" as the rule, ignoring the positive results of tests for an "Ascending numbers" hypothesis ("1-2-3" and "53-86-105"). Subject 400 used a limited number of tests (repeating "4-6-8" and "2-4-6") for 8 out of 11 trials and ignored the negative results of the other three trials ("8-10-12", "6-8-10", and "102-104-106"). The results of the latter three tests disconfirmed the final rule selection--"Ascending numbers".

In the Error condition, only two solvers openly referred to the possibility of error. In 6 out of 7 protocols, tests were

only repeated if they resulted in negative feedback. Subject 410 did not repeat tests, but used a combination of consistent and inconsistent tests. Such a strategy allowed the hypothesis choices to be narrowed down systematically. Subject 416 also narrowed down choices based on feedback from the tests, but only repeated one test. Subject 412 tried numerous hypotheses, reviewed the results of all tests three times and systematically eliminated ideas.

Non-solvers in the Error condition ignored data which disconfirmed their final rule selection and selected rules to tests which had already been ruled out by previous tests. Four non-solvers did make reference to the possibility of error. Subject 408 failed to retest a critical disconfirmatory error trial ("8-5-3") and selected "Ascending numbers less than 10" as the final rule. The correct rule was "Numbers less than 10". Subject 406 used repeated tests with different rules, but did not seem to understand that particular selections had been ruled out by previous tests. Subject 421 conducted 37 tests, repeated only four tests (two error trials), but ignored data from six tests which disconfirmed the selected final rule. The choice was based on the results of seven tests which confirmed it.

Discussion. The general results demonstrated that the think-aloud procedure did not adversely affect problem-solving behavior or performance. The difference in solving rates between the No Error and Error conditions was very similar to results found in previous experiments. Problem-solving behavior was differentially affected by condition as indicated by the increased number of tests conducted by subjects in the Error condition. In addition, the potential for successful solution was again decreased by the presence of error in the feedback.

A primary purpose of the present experiment was to explore problem-solving styles used in the current version of the Wason 2-4-6 task in a more detailed manner than had been previously attempted. The protocol analyses revealed several interesting phenomena which had not been captured earlier by routine analysis of rule selection, tests, expectations, and test results. First, non-solvers' tended to ignore test feedback which eliminated potential rules, but did not verbalize their reasoning for following such a strategy. It was felt that such subjects might not have understood the relationship of individual test results to all potential rules, but few expressed not understanding what they were supposed to do. Confusion over how to conduct and evaluate tests was also not evident in what subjects indicated they expected outcomes to be. Very few subjects chose the category "Unsure" when asked whether they expected their tests to fit the computer's rule and the clear majority usually expected tests to fit. Non-verbalization of strategy might also have been indicative of a lack of particular strategy for arriving at a possible conclusion.

Second, only a small number of subjects made reference to the possibility of error during testing, even when test results disconfirmed a selected rule. Non-verbalization of the possibility of error might have been indicative of failure to attend to or believe the warning message at the beginning of the task. Third, many solvers in both the No Error and Error conditions used a form of counterfactual reasoning by conducting tests which matched a selected hypothesis and simultaneously ruling out other potential hypotheses.

Experiment 9. Training vs. No Training, No Error vs. Error

Rationale. The purpose of the final experiment was to explore the use of a less abstract, but analogous problem-solving task as a method of conceptual training for the usual Wason 2-4-6 problem.

Subjects. Eighty (80) University of Central Florida undergraduate students were recruited for the study. All were given experimental credit for their participation.

Procedure. The experiment utilized the same basic procedures for the Wason 2-4-6 task under Untimed No Error and Error conditions that were followed in Experiments 5A, 6A, 7A, and 8. Subjects were randomly assigned to one of four conditions: (1) Untrained No Error; (2) Untrained Error; (3) Trained No Error; and (4) Trained Error. For the Trained No Error and Error conditions, subjects participated in a game with the experimenter in which they were asked to solve a murder mystery (see Figure 24) by collecting "evidence". Evidence collection was done by querying the experimenter and receiving feedback. Questions were answered according to a "data" sheet (see Figure 25) which contained information about the suspects and the crime scene. If a question was asked that was not relevant, the experimenter would indicate that the data was "not important." After all data had been collected, the subject was asked to name the murderer. Following the training task, all subjects were informed about the utility of using disconfirmation to rule out possible hypotheses and using repeated tests.

Results. In both the Untrained and Trained No Error conditions, 14 out of 20 (70.0%) subjects were able to solve the task, compared to 10 out of 21 (47.6%) subjects who were able to solve in the Untrained Error condition and 11 out of 19 (57.9%) subjects in the Trained Error condition. (See Figure 26.) The differences in solving rates among the conditions was not significant ($\chi^2(3, N = 80) = 3.024$, NS).

The mean number of tests conducted in the Untrained No Error condition was 6.5, compared to 5.7 tests conducted in the Trained No Error condition. The mean number of tests conducted in the Untrained Error condition was 12.0, compared to 13.1 tests conducted in the Trained Error condition. (See Figure 27.) A two-way ANOVA revealed a significant main effect for error, but not for training, in the number of tests conducted among the conditions ($F(1,76) = 20.145$, $p < .001$).

The mean percentage of total tests that were repeated in the Untrained No Error condition was 18.9%, compared to 24.4% in the Trained No Error condition. The mean percentage of total tests that were repeated in the Untrained Error condition was 29.3%,

Instructions

In this game you are playing the part of a detective. It is your job, as the detective, to solve a horrible murder. This can best be accomplished by asking a series of questions to the experimenter. The experimenter will act as the suspects and the crime lab assistant. All information you need to solve the crime, these three people can give you. The best way to solve the case is to ask questions of the three people. It is recommended that you establish the general facts from the crime-lab first, such as how the Dr. was murdered, what was in the missing file, and information regarding the cigarette butts. Then move onto the specific questioning of the suspects. Attempt to focus on their relationship with Dr. Falk, their smoking habits, and any information they knew regarding what was in the missing file. These questions will be used to determine which one was responsible for the murder of Dr. Falk. Move from general to specific. Remember, once you have ruled out one suspect, **do not** just assume that the other person was responsible for the murder of our beloved Dr. Falk. Use the provided sheet to keep track of your questions and answers as you are figuring out this crime. Now the story.

The Case of the Dead Professor

It was a pretty normal day, that bright day in January. But something was just not right. Earlier in the day a secretary had found the murdered body of Dr. Falk in his office. Of course, the police and local detectives were called in to handle the case. Our story begins with you, the local bigshot detective overlooking the office of the deceased.

As a seasoned veteran of the field you new exactly what you were looking for--clues. This is of course what detectives are supposed to do when they are examining the crime scene. As far as you could tell, the office appeared to be in normal order. However, there were several items worth noting: the ash tray contained cigarette butts of two different kinds of cigarettes, the file cabinet was open, and Dr. Falk's coat could not be found. Upon closer inspection of the file cabinet it was determined that a file regarding a particular grant was missing. This grant supported all of the graduate students that worked for Dr. Falk.

The graduate students would need to be questioned regarding their whereabouts at the time of the murder. Dr. Falk was in charge of two graduate students--Mark and Robert. It was known throughout the department that Dr. Falk was going to cancel the grant that supported the two graduate students. However, it was also rumored that one of the students would be supported under a new grant. Whatever this file truly held must have been motive enough to commit murder. The game has begun. Go ahead and begin your questioning.

Figure 24. Mystery scenario

Data on Mark

1. Third year graduate student.
2. Smokes Marlboro. (This is different from what the Dr. smokes.)
3. Has the file from Dr. Falk's cabinet. (It was loaned to him by the Dr. to go over and make sure that everything was in proper order.)
4. At the time of the murder, he was at his apartment studying. (His roommate is his witness along with several friends that called during the time in question.)
5. Was known to have his problems with Dr. Falk. (This mostly was in regards to the type of errands the doctor had Mark do.)
6. He was going to be supported under a new grant.
7. Did not murder Dr. Falk.

Data on Robert

1. Fourth year second year graduate student.
2. Does not smoke.
3. It was known that he had difficulty with the Dr. (This was in regards to the fact that he was not getting paid enough for the work he was doing.)
4. He has no alibi for the time of the murder.
5. He knew Dr. Falk was going to cancel the grant and not going to renew one that would support him.
6. Robert has the coat. (He used it to strangle and kill Dr. Falk.)

The Crime Lab's information

1. More than one person entered Dr. Falk's office the day of the murder besides Dr. Falk.
2. The two cigarette butts were of different brands. (Dr. Falk's brand was Camel.) (The other brand was Marlboro.)
3. The file contained important information about a grant that supported graduate students. (The important information in the file was that Dr. Falk was going to apply for a new grant to only support one student.) (The one student the new grant was going to support was Mark.)
4. Dr. Falk was murdered by strangulation. (It was done with a fabric of some sort.) (The coat was the tool used to murder Dr. Falk.)

Figure 25. Mystery data sheet

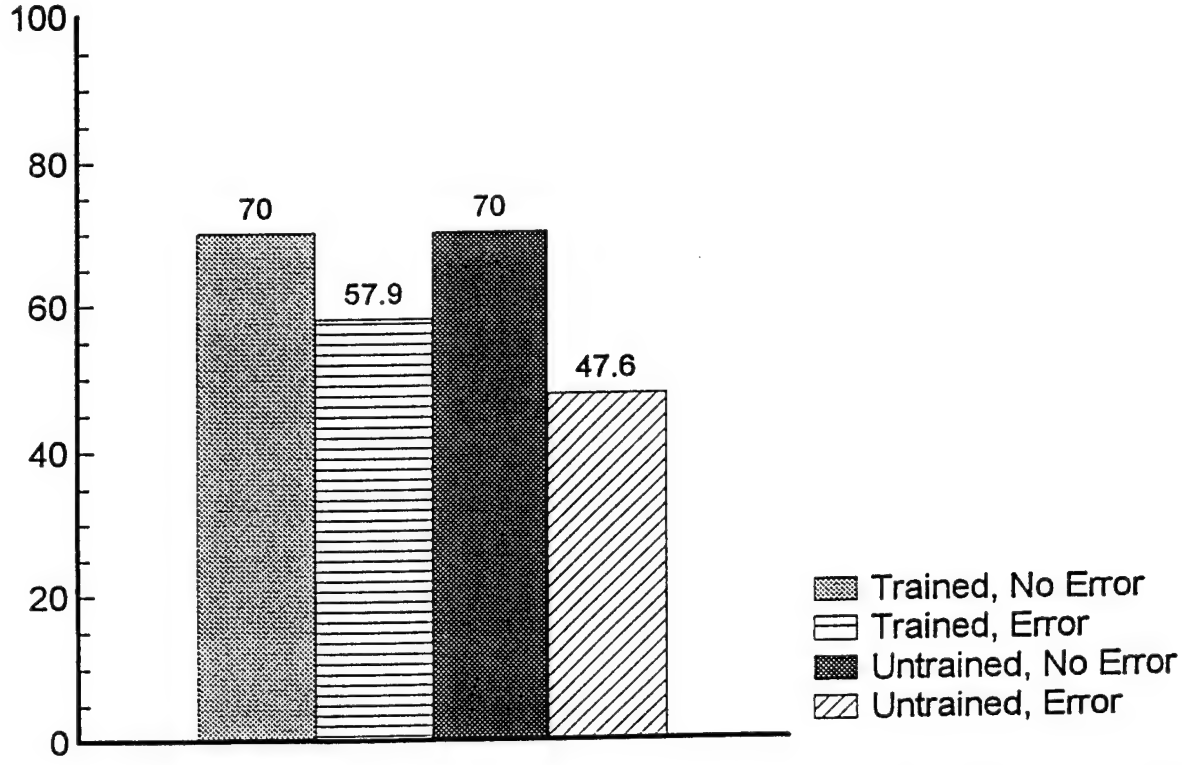


Figure 26. Experiment 9--Training vs. No Training, No Error vs. Error Percentage of Solvers

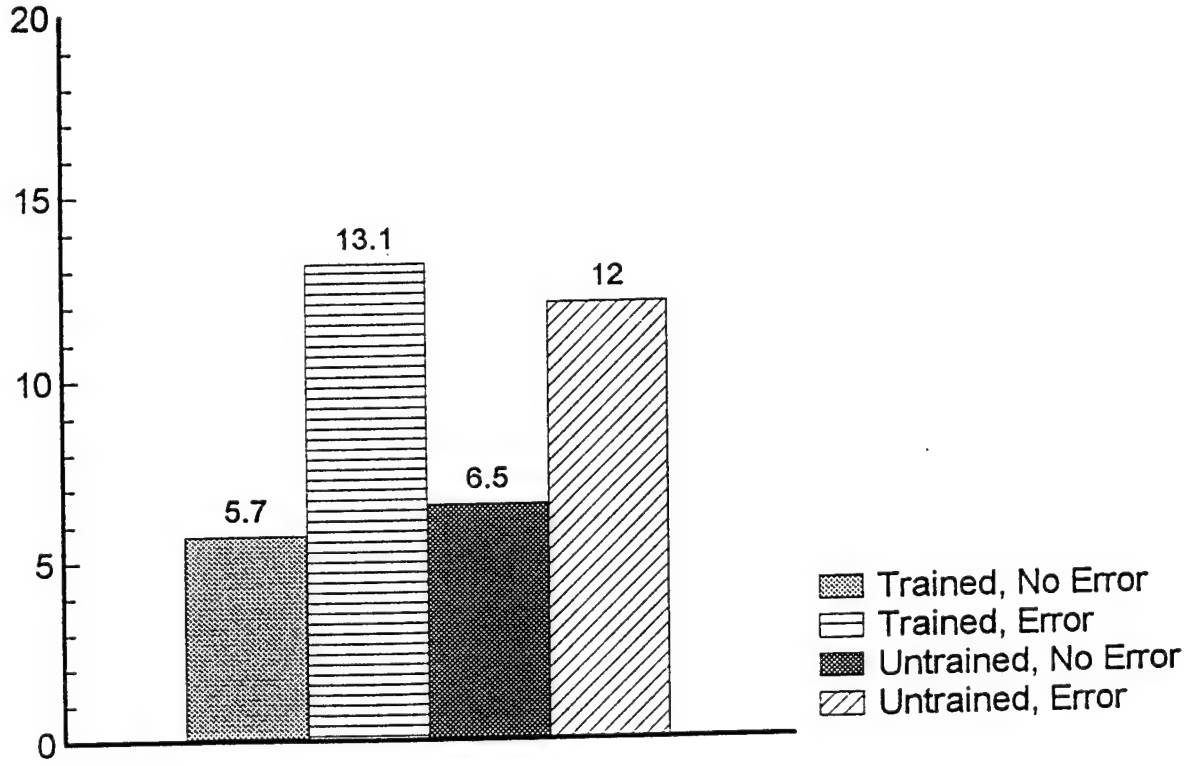


Figure 27. Experiment 9--Training vs. No Training, No Error vs. Error Mean Number of Tests Conducted

compared to 32.8% in the Trained Error condition. A two-way ANOVA again revealed a significant main effect for error, but not for training, in the percentage of repeated tests conducted among the conditions ($F(1,76) = 8.247$, $p = .005$). (See Figure 28.) There was also no significant effect of condition on the number of individuals who repeated tests compared to those who did not repeat tests ($\chi^2(3, N = 80) = 4.223$, NS). Nine out of 20 (45.0%) subjects in the Untrained No Error condition repeated tests, compared to 8 out of 20 (40.0%) subjects in the Trained No Error condition. Fourteen out of 21 (66.7%) Untrained Error condition subjects repeated tests, compared to 12 out of 19 (63.2%) Trained Error condition subjects.

A two-way ANOVA indicated no significant effects for training or error in the percentages of total tests attempted that subjects' expected would fit the computer's rule ($F(1,76) = .232$, NS; $F(1,76) = 3.72$, NS). Subjects in the Untrained No Error condition indicated that they expected 82.9% of their total attempted tests to fit the rule, while subjects in the Trained No Error condition expected 74.2% to fit. Subjects in the Untrained Error condition indicated that they expected 53.6% of their total attempted tests to fit the rule, while subjects in the Trained Error condition expected 70.6% to fit.

A significant main effect for error was found among the percentages of total tests attempted that subjects expected would not fit the computer's rule ($F(1,76) = 7.545$, $p = .008$). Subjects in the Untrained No Error condition indicated that they did not expect 9.4% of their total attempted tests to fit the rule, while subjects in the Trained No Error condition did not expect 3.5% to fit. Subjects in the Untrained Error condition indicated that they did not expect 31.1% of their total attempted tests to fit the rule, while subjects in the Trained Error condition did not expect 15.7% to fit. (See Figure 29.)

A two-way ANOVA indicated no significant main effects for training or error in the percentages of "unsure" responses ($F(1,76) = .935$, NS; $F(1,76) = .006$, NS). Subjects in the Untrained No Error condition indicated that they were unsure of 7.7% of their total attempted test outcomes, while subjects in the Trained No Error condition were unsure of 22.3% of the outcomes. Subjects in the Untrained Error condition indicated that they were unsure of 15.3% of their total attempted test outcomes, while subjects in the Trained Error condition were unsure of 13.7% of the outcomes.

There were no significant main effects for training or error found among the mean percentages of confirmatory test outcomes ($F(1,76) = 2.475$, NS; $F(1,76) = 1.868$, NS). In the Untrained No Error condition, 58.7% of the test outcomes were confirmatory, compared to 51.6% in the Trained No Error condition. In the Untrained Error condition, 52.8% of the test outcomes were

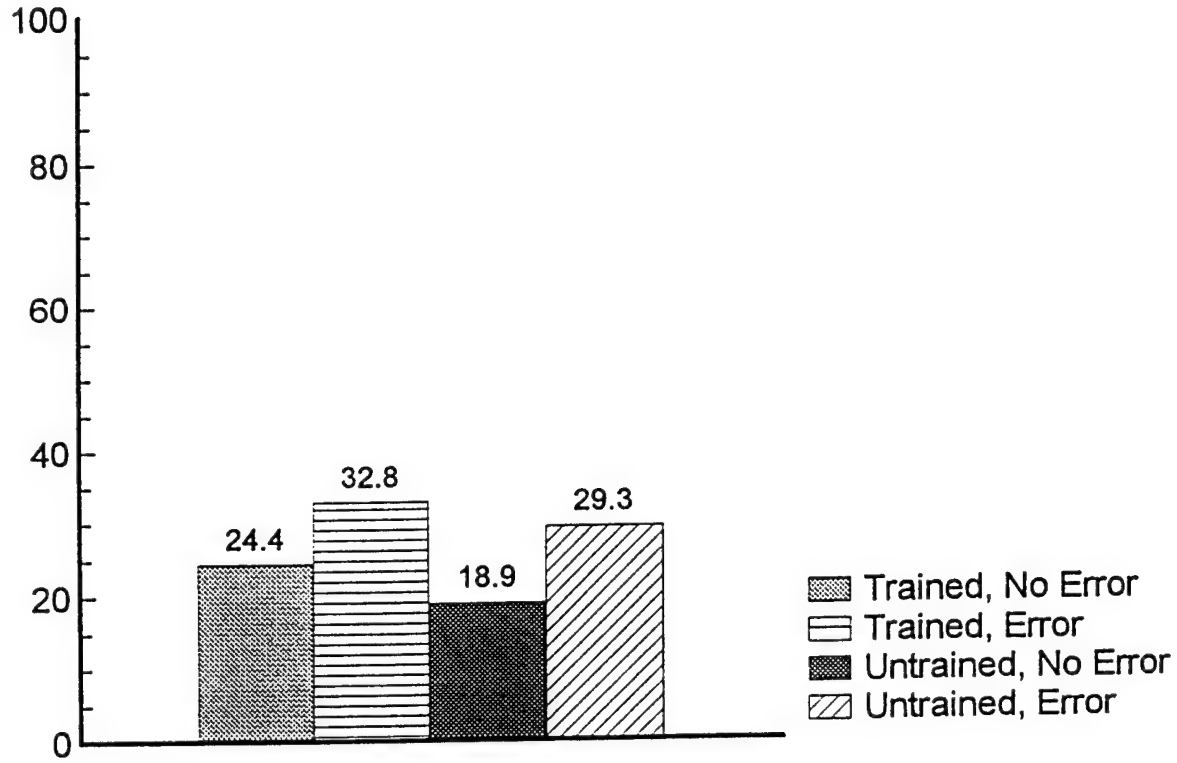


Figure 28. Experiment 9--Training vs. No Training, No Error vs. Error Mean Percentage of Total Tests Repeated

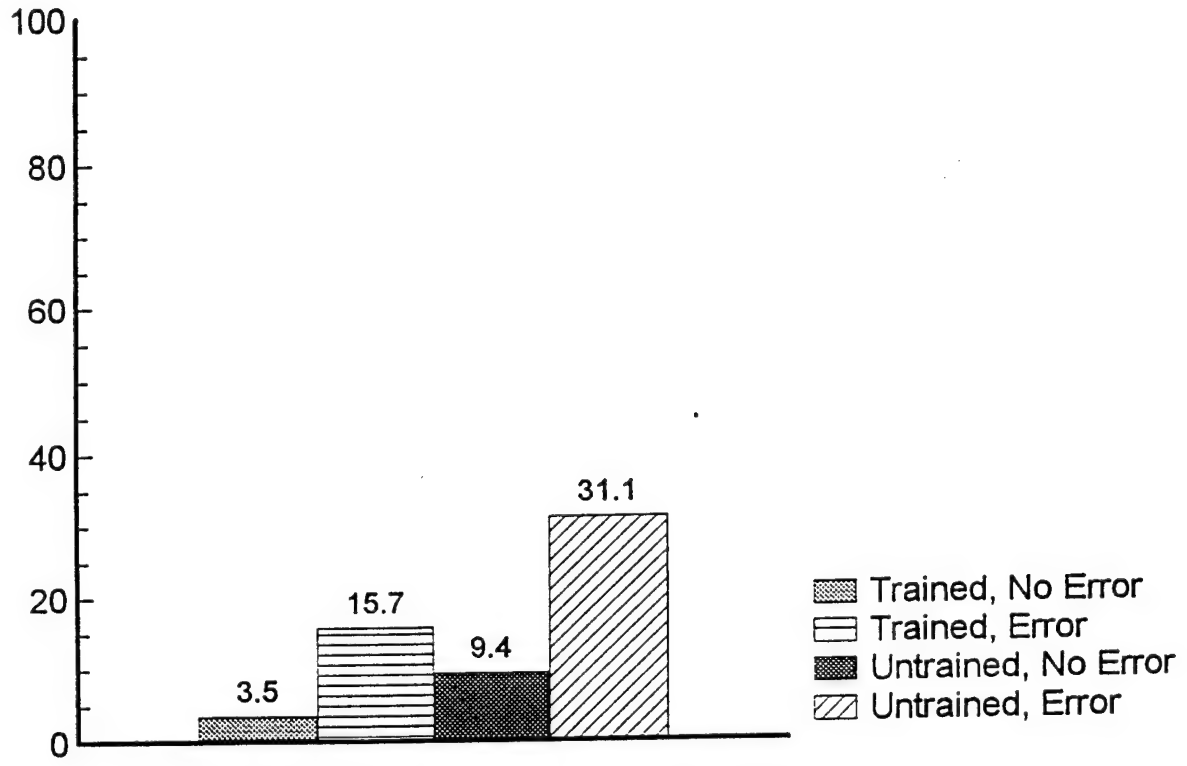


Figure 29. Experiment 9--Training vs. No Training, No Error vs. Error Percentage of Expected Negative Test Outcomes

confirmatory, compared to 41.7% in the Trained Error condition. There was a significant interaction between training and error found among the mean percentages of disconfirmatory test outcomes ($F(1,76) = 4.114$, $p = .046$). In the Untrained No Error condition, 33.5% of the test outcomes were disconfirmatory, compared to 26.1% in the Trained No Error condition. In the Untrained Error condition, 31.9% of the test outcomes were disconfirmatory, compared to 44.6% in the Trained Error condition.

Discussion. While the difference in solving rates between the No Error and Error conditions was not significant, it was in the expected direction. Again, the presence of data error degraded performance as had been found in the earlier experiments. However, the similarities in solving rates among the trained and untrained conditions failed to demonstrate that providing an analogous conceptual training task enhanced performance.

Though there was a significant increase found in the percentage of inconsistent tests attempted in the error conditions, it did not appear to be related to specific instructions to disconfirm. A higher percentage of inconsistent tests (31.1%) were conducted by subjects in the Untrained Error condition than were conducted by subjects in the Trained Error condition (15.7%). Furthermore, the Trained Error subjects utilized inconsistent tests at a rate almost identical to the rate found in the previous experiment. Thus, as previous studies (Tweney et al., 1980; Gorman and Gorman; 1984) had shown, instructing subjects in the utility of disconfirmation had little, if any, effect on their problem-solving behavior. Many subjects also repeated tests whether or not they were instructed to do so, though the number of subjects was fairly high for all conditions.

General Discussion and Implications

Performance. The majority of experiments repeatedly demonstrated that system failure degraded performance as measured by successful task solution in both the Wason and Tribbles problems. It was surprising and puzzling that the imposed time limitations in the Wason error conditions had differential effects in two experiments, which were conducted using samples from two different subject pools. It has been speculated that the differences in performance might be attributed in part to diferences in technological experience, but the study was not designed to assess this particular issue. It is strongly felt, however, that such factors should be taken into account when both designing and conducting future studies.

Heuristics. As exemplified by the Wason protocol analysis experiment, many subjects were able to consider several hypotheses and test results simultaneously to eliminate rules and arrive at the correct task solution. The strategy was not as effective in the error condition if critical tests were not repeated and/or if only tests which disconfirmed hypotheses were repeated. Error warnings appeared to alert the subject to potential problems, but for many the scope of potential error was considered to be unidimensional.

Implications. It is felt that the general finding of subjects' poor performance under system failure conditions is of particular importance in designing training programs for complex systems. If subjects are unable to modify their problem-solving behavior on such simple tasks to deal with data error from a single source, multiple tasks with numerous sources of potential error will pose an even greater problem. Furthermore, the strong detrimental effects found when system failure was combined with time constraints also must be considered. It is imperative that training programs that utilize simulation of complex systems require training under degraded mode and time constrained conditions. Such training should provide an opportunity for assessment of how trainees are reacting to the simulated reliability of the system and development of techniques for handling specific types of unreliable data.

References

Billings, C. E. (1991). Human-centered aircraft automation: A concept and guidelines. (Technical Memorandum 103885). Moffett Field, CA: National Aeronautics and Space Administration.

Bliss, J. P. (1993). The cry-wolf phenomenon and its effect on alarm responses. Unpublished doctoral dissertation, University of Central Florida, Orlando, FL.

Breznitz, S. (1983). Cry-wolf: The psychology of false alarms. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley, CA: University of California Press.

Caelli, T., & Porter, D. (1980). On difficulties in localizing ambulance sirens. Human Factors, 22, 719-724.

Doherty, M. E., & Tweney, R. D. (1988). The role of data and feedback error in inference and prediction. (In preparation). Alexandria, VA: Army Research Institute, Office of Basic Research.

Ericsson, K. A. & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Farris, H. H. & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. Memory and Cognition, 17(2), 221-232.

Gorman, M. E. (1986). How the possibility of error affects falsification on a task that models scientific problem-solving. British Journal of Psychology, 77, 85-96.

Gorman, M. E. & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory, and a control strategy on Wason's 2-4-6 task. Quarterly Journal of Experimental Psychology, 36A, 629-48.

Kantowitz, B. H. (1977). Ergonomics and the design of nuclear power plant control complexes. In T. W. Kvalseth, (Ed.), Arbeidsplass og miljobruk av ergonomiske data. Rtondheim, Norway: Tapir.

Kern, L. (1982). The effect of data error on inducing confirmatory inference strategies in scientific hypothesis testing. Unpublished doctoral dissertation, The Ohio State University, Columbus, OH.

Kerr, J. H. (1985). Auditory warnings in intensive care units and operating theatres. Ergonomics international, 85.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94, 211-228.

Mynatt, C. R., Doherty, M. E., & Tweney, R. E. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 29, 85-95.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. Quarterly Journal of Experimental Psychology, 30, 395-406.

Pate-Cornell, M. E. (1986). Warning systems in risk management. Risk Analysis, 6(2), 223-234.

Platt, J. R. (1964). Strong inference. Science, 146, 347-353.

Tukey, D. D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. Quarterly Journal of Experimental Psychology, 38A, 5-33.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. Quarterly Jounral of Experimental Psychology, 32, 109-23.

Tweney, R. D. (1985). Faraday's discovery of induction: A cognitive approach. In D. Gooding & F. James (Eds.), Faraday rediscovered (pp. 159-206). London: MacMillan.

Walker, B. J. (1985). Variations in problem solving styles in the Wason 2-4-6 rule discovery task. In D. R. Moates & R. Butrick (Eds.), Proceedings Inference OUIC 86 (pp. 280-289). Athens, OH: Ohio University.

Walker, B. J. (1987). A comparison of the psychological effects of the possibility of error and actual error on hypothesis testing. Unpublished doctoral dissertation, Bowling Green State University, Bowling Green, OH.

Walker, B. J. & Harper, D. R. (1989). Effects of data error on problem-solving heuristics. (AFOSR Contract No. F49620-85-C-0013). Dayton, OH: Universal Energy Systems.

Walker, B. J. & Harper, D. R. (1990). Decision-making under system failure conditions. (AFOSR Contract No. F49620-88-C-0053). Dayton, OH: Universal Energy Systems.

Walker, B. J. & Tweney, R. D. (1986, May). Protocol analysis of the Wason 2-4-6 rule discovery task. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-40.

York, K. M., Doherty, M. E., & Kamouri, J. (1987). The influence of cue unreliability on judgment in a multiple cue probability learning task. Organizational Behavior and Human Desision Processes, 39, 303-317.